DOCUMENT RESUME

ED 302 693 CE 051 719

TITLE Corrections Officer Validation Report and Technical Adequacy Report. Standards and Training for Corrections Program.
INSTITUTION California State Board of Corrections, Sacramento.
PUB DATE 87
NOTE 130p.; For related documents, see CE 051 717-725, 729-735, and 739-744.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Adult Education; Correctional Rehabilitation; Criterion Referenced Tests; Employment Practices; Entry Workers; Evaluation Criteria; *Occupational Tests; *Personnel Selection; Postsecondary Education; Predictive Validity; Predictor Variables; Research Methodology; Sample Size; Sampling; State Standards; *Statewide Planning; Test Construction; Testing; *Test Validity
IDENTIFIERS *California; *Correctional Officers

ABSTRACT

In response to a mandate from the California State Legislature, the California Board of Corrections conducted a three-phase research study in preparation for the revision and standardization of the current state standards for hiring and training corrections personnel. The primary focus of the second phase of the project was on the development and validation of entry-level selection procedures for these positions. This set of technical reports deals with the process of developing and validating the criteria on which the selection procedures developed were based. The validation report covers the following topics: the project goals and the research strategies used to achieve them; the principal results of the job analysis; the predictor and criterion development processes; data collection procedures used; the criterion, predictor, and validity and fairness analyses used in developing the selection battery; and recommendations pertaining to administering and scoring the test. The technical adequacy report discusses the validity of the testing program in view of the fact that, although the original project plans called for testing a minimum of 500 probation officers, the final validation sample ranged from 250 to 339 officers depending on the analysis being conducted. (YLB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302693

CORRECTIONS OFFICER VALIDATION REPORT AND TECHNICAL ADEQUACY REPORT

STANDARDS AND TRAINING FOR CORRECTIONS PROGRAM

CE051719

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C.W. Page

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

STANDARDS AND TRAINING
FOR CORRECTIONS PROGRAM

STC

BOARD OF CORRECTIONS

VALIDATION REPORT

CORRECTIONS OFFICER

Copyright 1987 California Board of Corrections

In preparing the Corrections Officer Examination for the California Board of Corrections, Personnel Decisions, Inc. has taken into account the following documents concerning selection testing:

- The American Psychological Association, Standards for Educational and Psychological Tests, (1974).
- The APA Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures, 2nd edition, (1980).
- The Equal Employment Opportunity Commission, Uniform Guidelines on Employee Selection Procedures, (1978).

CONTENTS

Page

List of Tables ii
List of Figures iv

Chapter 1 INTRODUCTION

Project Background..... 1
Participating Departments..... 2
Research Strategy and Rationale.....2
Report Overview.....4

Chapter 2 SUMMARY OF THE JOB ANALYSIS 5

Chapter 3 PREDICTOR DEVELOPMENT

Overview and Rationale.....7
Linking KSAs to Job Tasks.....9
Grouping Task Items into Task Categories.....13
Task Category to Worker Characteristic Ratings.....13
Setting Linkage Criteria.....18
Linking Tests to Worker Characteristics.....18
Linking Tests to Task Categories.....29
Predictor Development.....32
Test construction.....32
Pretesting the preliminary tests.....36
Item analysis and revision.....38
Special revision methods.....40
Final Test Battery.....41
Noncognitive Tests.....41
Sample Items Booklet.....42

Chapter 4 CRITERION DEVELOPMENT

Collecting Critical Incidents.....43
Revising and Retranslating the Critical Incidents.....44
Developing Behavioral Summary Scales.....48
Other Ratings Scales.....48
Worker characteristic ratings.....48
Overall effectiveness.....50

Chapter 5 DATA COLLECTION

Predictor Data Collection.....51
Criterion Data Collection.....52
Final Validation Sample.....53
Missing name.....55

CONTENTS (Chapter 5 continued)

Page

Missing ratings.....55
Low tenure.....55
Inaccurate ratings.....55
Demographics of Final Validation Sample.....55
Technical Adequacy of Sample.....59

Chapter 6 ANALYSES AND RESULTS

Criterion Analyses.....61
Data collection and cleaning.....61
Developing an overall criterion.....63
Search for additional criteria.....66
Criterion descriptive statistics.....69
Predictor Analyses.....73
Data collection and cleaning.....73
Scale revisions and final psychometrics.....73
Descriptive statistics.....74
Validity and Fairness Analyses.....81
Adverse impact and test fairness.....81
Validity and fairness of the individual tests.....82
Developing a Final Selection Battery.....85
Differential Predictability of Specific Criteria.....90

Chapter 7 DISCUSSION AND RECOMMENDATIONS

Exam Format and Content.....92
Using the Corrections Officer Examination.....94
Exam Security.....94
Scoring and Data Collection.....97
Score Interpretation and Cutoff Scores.....99

Chapter 8 TYPE I FACILITIES SURVEY
Corrections Officer Selection and Training.....100

References 101

TABLES

		Pages
3.1	Definitions of 24 Corrections Officer Worker Characteristics	10-12
3.2	31 Task Categories for Corrections Officers	14-15
3.3	Average Importance Ratings of 24 Worker Characteristics for 31 Task Categories	16
3.4	Reliabilities and Ratings Summary for Corrections Officers	17
3.5	Average Usefulness Ratings of 72 Tests as Measures of 16 Before-Hire Worker Characteristics	20-28
3.6	Reliabilities and Ratings Summary of Test to Worker Characteristics Ratings	30
3.7	Linkage Criteria and Summary for Test to Worker Characteristic and Worker Characteristic to Task Category Ratings	31
3.8	Cognitive and Noncognitive Tests for the Experimental Selection Battery	33-34
3.9	Pretest Results and Item Statistics for the Experimental Selection Battery	37
4.1	Corrections Officer Performance Dimensions Used During Retranslation	46-47
5.1	Composition of the Final Validation Sample	54
5.2	Gender and Ethnic Status of Final Validation Sample	56
5.3	Education and Tenure of Final Validation Sample	57
5.4	Relationship of Incumbents to Supervisors in Final Validation Sample	58
6.1	Descriptive Statistics for 32 Job Performance Ratings Scales	62
6.2	Correlation of 32 Job Performance Rating Scales with the Overall Criterion	65

TABLES (continued)

		Pages
6.3	Factor Descriptions of Five Specific Job Performance Criteria	67
6.4	Descriptive Statistics and Intercorrelations for Five Factor Score Criteria	70
6.5	Descriptive Statistics for the Overall and Specific Job Performance Criteria	71
6.6	Average Overall Criterion Scores by Supervisor and Incumbent Gender	72
6.7	Average Overall Criterion Scores by Supervisor and Incumbent Ethnic Status	75
6.8	Final Item Statistics for the Experimental Selection Battery	76
6.9	Descriptive Statistics on 15 Individual Tests for the Final Validation Sample	77
6.10	Intercorrelations Among Individual Tests	78
6.11	Descriptive Statistics on Individual Tests by Incumbent Gender	79
6.12	Descriptive Statistics on 15 Individual Tests by Incumbent Ethnic Status	80
6.13	Correlations of 15 Individual Tests with Overall Job Performance	83
6.14	Source of Items Selected to Augment the Final Selection Battery	87
6.15	Gender and Ethnic Group Statistics for Final Selection Battery Scores	88
6.16	Correlations of Individual and Final Predictors with the Five Factor Criterion Measures	91
7.1	Description of the Noncognitive Tests Contributing Items to the Corrections Officer Examination	95-98
7.2	Corrections Officer Examination Score Distribution	98

FIGURES

		Page
3.1	Hypothetical Example of the Relationships Among Tests, KSAs, and Job Tasks	8
3.2	Test Development Plan for the Numerical Ability Test	35
4.1	Sample Performance Incident Form	45
4.2	Sample Behavioral Summary Scale	49
6.1	Factor Loadings of 31 Ratings Scales on Five Specific Criteria Factors	68

CHAPTER ONE: INTRODUCTION

The research, analysis, and results described in this report represent the second phase of a three-phase research study sponsored by the Standards and Training for Corrections Program (STC) of the California Board of Corrections. This introduction provides general information about the project and describes the strategy used to conduct the research.

Project Background

The California Board of Corrections had been charged by the State Legislature to undertake several previous studies of the probation (1964) and correctional systems (1971 & 1979). These studies found widely varying hiring standards and training practices from agency to agency -- even where printed job descriptions were quite similar. Additionally, many agencies were finding themselves in court faced with the task of defending their hiring procedures while incurring the considerable expense associated with such litigation and facing substantial delays in filling vacancies.

To address these problems, the Board of Corrections made two recommendations to the Legislature. The Board recommended that:

1. A state agency assume responsibility for an analysis and synthesis of job titles and job descriptions and for validation studies of selection standards for corrections personnel.

2. A commission be established, modeled after POST, specifically for standards setting and training for corrections personnel.

In response to the Board's recommendations, the California Probation, Parole, and Correctional Association (CPPCA) proposed Senate Bill 924 which became effective July 1, 1980. SB 924 created the Standards and Training for Corrections Program (STC) to establish minimum standards for the selection and training of local corrections and probation personnel, to certify training courses, and to administer the corrections training fund to help agencies comply with the training standards. Funds for this program are derived from penalty assessments on criminal offenses. Participation by individual agencies in the STC program is voluntary. Only those agencies accepting state funds through STC are required to adhere to the selection and training standards.

Subsequently, Assembly Bill 3296 was passed. This bill mandated the Board of Corrections to contract (among other things) for research on validated selection standards for corrections and probation personnel. This mandate provided the impetus for the research efforts described in this report and in related documents.

Participating Departments

County sheriff's or probation departments participating in the STC program were eligible to participate in this project. The 37 eligible sheriff's and 52 eligible probation departments were invited to attend one of three briefings where the goals and procedures of the study were presented. Approximately 20 sheriff's and 21 probation departments attended these briefings.

Departments unable to send representatives to the briefings were mailed copies of the briefing materials and participation agreement forms. Initially, 23 sheriff's departments and 30 probation departments (representing 62% and 58% of eligible departments) elected to participate in the study. The number varied from time to time as departments withdrew or were added to the study.

Research Strategy and Rationale

The research program has consisted of three phases, each corresponding to one of three objectives of the Standards and Training for Corrections Program. These three objectives are:

Phase 1 -- a statewide job analysis of three entry-level positions in county agencies: corrections officer, probation officer, and juvenile counselor.

Phase 2 -- the development and validation of entry-level selection procedures for these positions.

Phase 3 -- the development of training standards and curricula for these positions.

This report focuses primarily upon the second phase of the research project, namely the development and validation of entry-level selection procedures. It is important to note that this report concerns the corrections officer position only. Separate reports describe the parallel efforts undertaken for the probation officer and juvenile counselor positions also included in this project. Of necessity, this report refers frequently to procedures and analyses conducted during the job analysis phase of the project, but is not to be a complete description of that research which is described fully elsewhere.[1]

[1]Grabow, K.M., Sevy, B.A., Houston, J.S. Statewide Job Analysis of Three Entry-level Corrections Positions, Personnel Decisions, Inc., 1983. (Available from the Board of Corrections.)

In addition to this report, information about the methods and results of the phase two research for corrections officers may be found in the User's Manual as well as in the Candidate Information Booklet and, to a lesser degree, in the test booklet. To maintain test security access to the test booklets is strictly controlled and closely monitored.

A concurrent, criterion-related strategy was used to guide the development and validation of selection procedures for the entry-level corrections officer position. The basic goal of the selections standards phase of the project was to devise a method of accurately predicting the job performance of entry-level corrections officers. The criterion-related approach to meeting this goal consisted of developing measures of ability and personal characteristics (predictors) which could be shown to be statistically and meaningfully related to job performance. In addition, the predictors had to be fair and not impose heavy administrative burdens to implement. A concurrent study is one in which predictor and criterion data are collected from job incumbents.

Professional standards (APA, 1980) suggest that a criterion-related approach is preferable whenever it is feasible and provide guidelines for determining the feasibility of such an approach. The first and, for this study, the most important guideline, concerns the stability of the job. Criterion-related studies are not feasible if the position being studied is in a period of rapid flux. Professional standards are most directly concerned with the stability of a position over time but, due to the variety of departments providing incumbent corrections officers to the study, stability of the position across agencies must also be considered. With respect to stability across agencies, job analysis data were analyzed to determine the extent to which the corrections officer job (as described by incumbents and supervisors) varied as a function of several parameters (e.g., county size, type of facility, job shift). The results of these analyses clearly demonstrated the equivalence of the job from department to department, thus indicating that the stability of the job across agencies cannot be questioned.

The question of the stability of the Corrections Officer position over time cannot be as directly addressed, but it seems clear that the position is not in a "period of rapid evolution". Nevertheless, the data in this report reflect only a two and one half year "snapshot" of the corrections officer position. Should the job undergo significant changes in the future, it will be incumbent upon users of the selection procedures developed in this project to ensure that the procedures remain valid, fair, and useful.

Other guidelines relevant to determining the feasibility of a criterion-related strategy concern the potential for developing a relevant, reliable measure of job performance (criterion); and for collecting predictor and criterion data from a sample of persons that is "reasonably representative of the populations of peoples and jobs to which the results are to be generalized" (APA, 1980).

It is also important that this sample be large enough to allow for reliable detection of meaningful relationships between the predictors and criteria. The research described in this report fully met each of these guidelines. Therefore, a criterion-related approach to the development and validation of entry-level selection standards for corrections officers was completely feasible.

Report Overview

The remainder of this report describes the methods used to develop and validate entry-level selection procedures. This description begins in Chapter Two with a summary of the job analysis results that formed the foundation for predictor and criterion development. Development of the predictors and criteria are described separately in Chapters Three and Four. Chapter Five describes the methods used to collect predictor and criterion data from a sample of approximately 300 corrections officer incumbents and their supervisors. Chapter Six reports the methods used to analyze the predictor and criterion data and documents the results of those analyses. Finally, Chapter Seven reviews the results with an eye towards their interpretation and the implications they have for use of the Corrections Officer Examination as part of the pre-employment selection process of participating agencies.

CHAPTER TWO: SUMMARY OF THE JOB ANALYSIS

The job analysis of the corrections officer position began with a review of job descriptions, training and orientation materials, and performance evaluation forms supplied by participating sheriff's departments. Document review was followed by site visits to eight counties. During the site visits, detention facilities were toured and corrections officer incumbents and supervisors interviewed individually and in groups to determine the major duties and tasks involved in the corrections officer position and the important requirements of the job.

Information gathered from the document review and site visits was used to develop a preliminary version of a job analysis questionnaire (JAQ). The preliminary questionnaire was reviewed by STC staff and by supervisors, training officers, and others from a sample of participating departments. Based on the reviewers' comments, the questionnaire was revised and then pilot tested on a sample of corrections officer incumbents and supervisors. During the pilot test, the incumbents and supervisors completed the preliminary questionnaire and provided feedback about its clarity and comprehensiveness.

Based upon the data collected in the pilot test, final revisions of the job analysis questionnaire were made. Two forms of this questionnaire were developed, one for corrections officer incumbents and the other for their supervisors. Each questionnaire consisted of four major sections: (1) a general section, collecting information about the person completing the questionnaire; (2) a task rating section where incumbents rated how often they performed each of 194 job tasks and supervisors rated the importance of each of those tasks; (3) an equipment rating section in which each of 47 pieces of equipment were rated with respect to their frequency of use on the job; and (4) a worker characteristics section in which each of 105 knowledges, skills, abilities, and personal characteristics were rated in terms of their usefulness for successful job performance.

The final job analysis questionnaires were administered to a random sample of 578 experienced incumbents (those on the job at least six months) and most of the supervisors in the participating counties (205 supervisors). In addition, 71 incumbents and 39 supervisors completed the questionnaire a second time in order to determine its test-retest reliability.

Analysis of the completed questionnaires resulted in a statewide definition of the corrections officer "core job", as well as specific definitions of how the job is performed in each

participating county. The core job was defined, based upon the task ratings from all participating counties, as consisting of all job tasks listed in the job analysis questionnaire which:

1. were performed by half or more of the incumbents across most counties,

OR

2. were rated as having medium or higher importance by the supervisors in most counties.

Both criteria were computed using a procedure that weighted all departments equally. This was done to ensure that the results would be applicable to all participating departments and so that the results from large departments did not completely dominate the study. Job tasks not meeting one of the above criteria were excluded from the core job definition -- as were those tasks with medium or higher importance but which were rarely performed in most departments.

Apart from defining the core job for the corrections officer position, the job analysis data showed that the job analysis questionnaire was highly reliable over time; hat the corrections officer job did not vary from county to county. This latter finding was very important as it permitted the development of a uniform set of selection procedures to be applied to agencies throughout the State of California.

For a more detailed description of the research methodology involved in the job analysis, see the Job Analysis Technical Report, December, 1983. (Available from the Board of Corrections.)

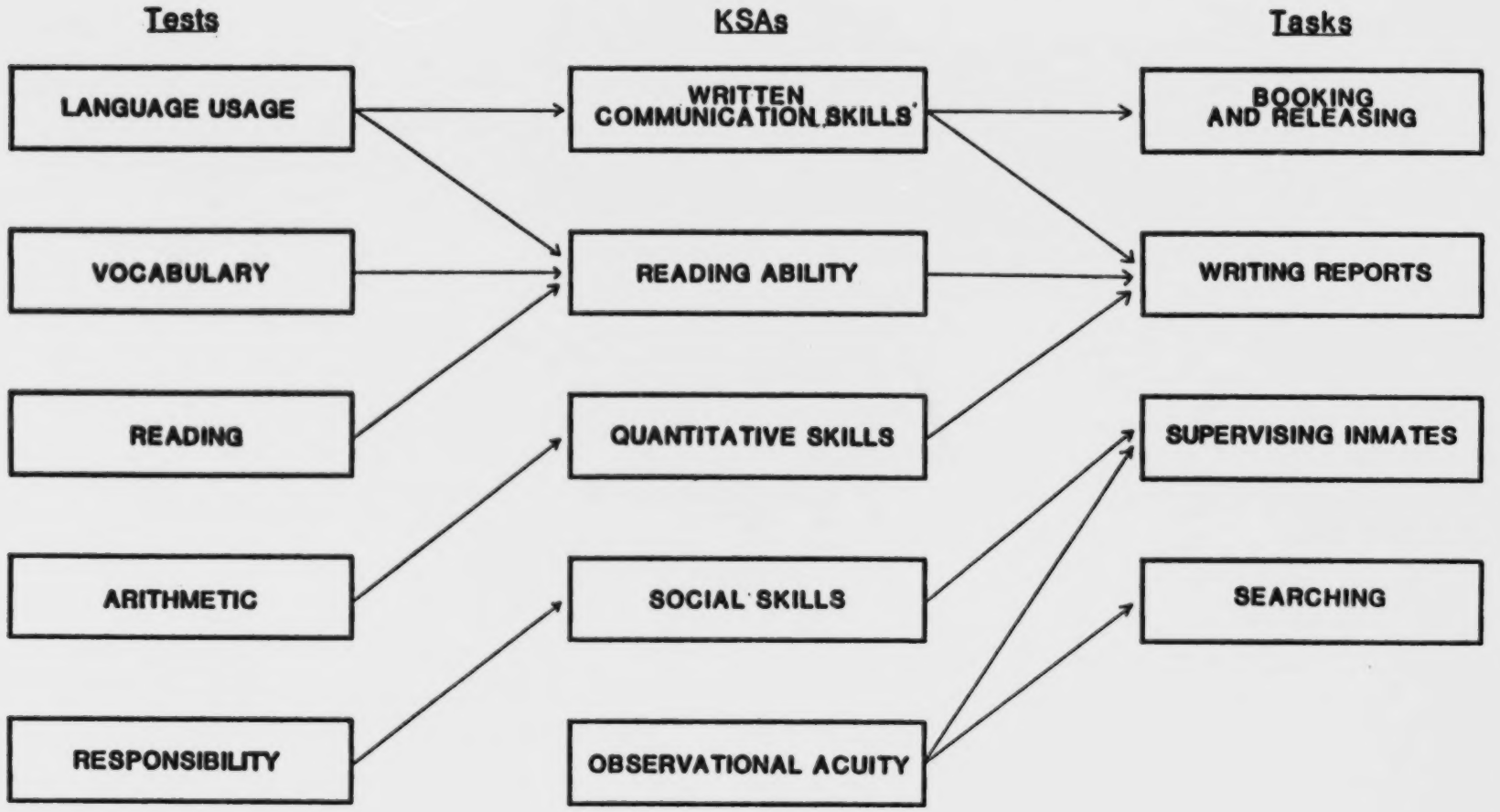
CHAPTER THREE: PREDICTOR DEVELOPMENT

Overview and Rationale

The end goal of predictor development was to arrive at a set of tests (experimental selection battery) to be administered to a sample of corrections officer incumbents (validation sample). These tests would comprise the building blocks from which the Corrections Officer Examination could be constructed. Because the tests included in the experimental selection battery would form the basis of the final, implemented examination, it was very important that the tests be highly indicative of candidates' ability to perform all important job tasks. It was necessary that the tests predict performance on the tasks that make up the job. However, tests do not measure task performance directly. What tests do measure are a person's knowledges (e.g., knowledge of a facility's rules and regulations), skills (e.g., skill in adding or subtracting), abilities (e.g., ability to remain alert and aware), or personal characteristics (e.g., integrity). In order to link the tests to the job tasks, it was necessary to use the knowledges, skills, and abilities (KSAs) from the job analysis as intermediaries. Tests measure KSAs, not job tasks. Once it could be shown which KSAs are essential to the performance of which tasks, tests could be selected that measure the KSAs and, therefore, that predict task performance. Figure 3.1 shows the hypothetical relationships among a set of tests, KSAs, and tasks. The arrows show which tests are useful measures of which KSA and which KSAs are required for the successful performance of which tasks. Successful report writing, for example, requires three KSAs (Written Communications Skills, Reading Ability, and Quantitative Skills). Four tests are useful measures of one or more of these three KSAs. Because those four tests are useful measures of the set of KSAs necessary for successful report writing, they can be used to predict report writing performance. In this project, the determination of which KSAs are important for specific tasks was made by corrections officer supervisors who rated the importance of every KSA for the successful performance of every job task. The usefulness of each of 72 tests as a measure of the KSAs was rated by PDI staff psychologists experienced in the use of tests in making selection decisions.

Chapter 3 describes the methods used to link tasks and tests to KSAs. Also in Chapter 3 is a description of how several new tests were drafted, pretested, and refined for inclusion in the experimental selection battery.

Figure 3.1 Hypothetical Example of the Relationships Among Tests, KSAs, and Job Tasks



Linking KSAs to Job Tasks

As was noted, one aspect of predictor development involved the linking of KSAs (directly measurable) to job tasks (ultimate criterion). Linking was accomplished by having job experts (corrections officer supervisors) rate the importance of each knowledge, skill, or ability for successful performance of the job tasks. Because there were 194 job tasks and 105 KSA items in the job analysis questionnaire, a reduction in the number of elements to be rated was necessary if the task was to be administratively feasible.

Reducing the number of KSAs to be rated was accomplished by combining them into broader groupings. First, PDI staff members rationally clustered the 105 individual KSAs into groups representing common characteristics (see Table 3.1). For example, several KSA items concerning aspects of facility-specific knowledge were grouped together. In this way, the 105 KSAs were combined into 24 worker characteristic clusters. This achieved a four-fold reduction in the number of KSAs to be rated, and at the same time, preserved the variety of skills and abilities important for successful job performance. The resulting worker characteristic clusters were reviewed by other PDI staff-members and then "confirmed" via corroborative factor analysis.

Confirmation consisted of two separate oblique factorings of the worker characteristic data: one using incumbents' ratings of the usefulness of each KSA for successful job performance, the other using similar ratings from supervisors. The corroborative factor analysis began with the assignment of every KSA to a target factor (i.e., one of the 24 worker characteristic clusters). Then, based upon the intercorrelations of the KSAs, the appropriate matrix algebra equations were solved yielding a new matrix that showed the relationship between each KSA and the 24 target factors (Overall & Klett, 1971). This matrix was examined to corroborate the rational clustering of the KSAs. The clustering was corroborated when the KSAs correlated more highly with their target factor than they did with any of the other factors. Generally the worker characteristics were corroborated, but the results suggested that a few KSAs were more strongly related to other clusters. So long as the relocations made sense, KSAs were relocated and their correlations recomputed to corroborate the new structure.

As noted, corroboration was done separately using the JAQ data of incumbents and supervisors. It was found that the KSA items migrating, for example, from worker characteristic "A" to worker characteristic "F", when incumbents' data were analyzed had a high probability of making the same migration when supervisors' data were analyzed. This indicated that the rational clustering of KSAs into worker characteristics and subsequent minor revisions provided reliable results.

Table 3.1

Definitions of 24 Corrections Officer Worker Characteristics

A. Facility-specific Knowledge Knowledge of a facility's: general rules and procedures; inmate discipline, search/shakedown, and emergency procedures; physical layout, shift schedules, security and communication systems; inmate classification system; administrative and report-writing forms and procedures.

B. Knowledge of the Legal System Knowledge of the history and philosophy of the criminal justice system, of court procedures and forms, and of state and federal laws pertaining to: criminal law enforcement, arrest, search, and seizure; local detention and holding facilities (e.g., minimum jail standards); and to collection, preservation, and identification of evidence.

C. Psychological Training Knowledge of: indicators of emotional disorders; human development; crisis intervention procedures; group leadership techniques; principles and practices of counseling in a custodial/rehabilitative setting.

D. Street Knowledge Knowledge of: inmate attitudes and behaviors; narcotics; gangs and gang behavior; and street terminology.

E. First Aid Knowledge of and skill in: basic first aid and CPR principles, procedures, and techniques.

F. Using Firearms Knowledge of and skill in the use of firearms.

G. Operating Communications Equipment Skill in the use of radio, intercom, switchboard, and other oral communications equipment.

H. Oral Communication Skill in: giving clear verbal instructions/directions, testifying, and other oral communication; understanding and following oral instructions from others.

I. Reading Skill in reading to: understand facility forms, reports, and manuals; understand/interpret laws, statutory codes, and court and other legal documents.

J. Written Communication Knowledge of: grammar and spelling; skill in completing forms and writing understandable and legible reports.

K. Quantitative Skill Skill in adding and subtracting.

Table 3.1 (continued)

L. Observation Acuity Skill in: accurately observing, recalling, and documenting behavior, visual details (such as identification marks), and other events; discerning and assessing security-risk factors, signs of potential problems, and subtle changes in behavior patterns; remaining alert and aware, even during long periods of inactivity.

M. Driving Skill in driving automobiles, vans, trucks, buses.

N. Self-defense/Physical Restraint of Others Knowledge of and skill in: application of self-defense, physical restraint, and riot control techniques and procedures; ability to apply the appropriate "use of force".

O. Social Skills Ability to maintain: discipline without causing unnecessary tension; effectiveness in working with others and interacting with diverse individuals and groups; consistency and fairness in dealing with others; insight, sensitivity, and concern for others' feelings; tolerance for differences in values and beliefs; sense of humor.

P. Motivation/Dependability Dependability and trustworthiness (e.g., coming to work on time, not abusing sick leave, assuming a fair share of the workload); willingness to put in extra time and effort; persistence in seeking information; integrity in adhering to professional ethics; commitment to a facility's goals and objectives.

Q. Stress Tolerance Tolerance for the stress of working in a potentially dangerous environment; ability to maintain self-control when confronted with verbal harassment or threats; ability to stay calm and respond with composure in emergencies.

R. Dominance/Self-esteem Assertiveness, self-assurance, and decisiveness; ability to influence and confront others, to enforce rules, to interact confidently with high-status individuals (e.g., judges), to make decisions where there is no standard procedure, and to delegate responsibilities and duties properly.

S. Non-autonomy Ability to work under close supervision; tolerance of highly structured/regimented work environment; willingness to obey authority.

TABLE 3.1(continued)

T. **Flexibility** Resourcefulness in thinking of alternative approaches (e.g., alternative treatment techniques, referral sources, etc.); ability to adjust to new and unfamiliar situations/duties; resilience in coping with setbacks.

U. **Tolerance of Negative Aspects of the Job** Tolerance of physically unpleasant work environments; ability to perform repetitive tasks effectively or tolerate long periods of inactivity.

V. **Searching** Skill in conducting pat- and strip-searches and in searching cells/dorms, vehicles, and other areas.

W. **Vision and Hearing Abilities** Daytime, nighttime, peripheral, and color vision; depth perception; and hearing acuity to be able to detect subtle changes in noise levels and quiet sounds such as whispering or scuffling.

X. **Physical Abilities** Upper body and leg strength; whole body coordination; manual dexterity and eye/hand coordination; endurance/stamina; and agility/flexibility.

Grouping Task Items Into Task Categories

In the job analysis questionnaire, the 194 individual task items had been grouped into 10 broad categories of job responsibility (e.g. Searching and Securing). For the same reason that the KSA items were grouped into a fewer number of worker characteristics, it was necessary to group the task items into a fewer number of task categories. This reduction was accomplished by rewriting the individual task items in each general JAQ category into one to five intermediate categories, each consisting of task items related to similar behaviors or responsibilities (e.g., obtaining information for booking purposes and filling out booking forms; see Table 3.2). The resulting 31 task categories (a six-fold reduction in number) were reviewed for completeness and used to link job tasks to worker characteristics.

Task Category to Worker Characteristic Ratings

Twenty-one corrections officer supervisors from five counties rated the importance of 24 worker characteristics for the successful performance of each of the 31 task categories (approximately 750 ratings per supervisor). On average, the raters were responsible for supervising 9 corrections officers and had been supervisors for an average of 4.7 years. They also had themselves served as corrections officers for an average of 3.8 years. Ratings were made using a 5-point scale, indicating the importance of each worker characteristic for every task category (1 = not important, 2 = low, 3 = medium, 4 = high, and 5 = very high importance). Table 3.3 contains the average importance rating of each worker characteristic for every task category.

The reliability of the ratings was estimated using the intraclass correlation approach described by Shrout and Fleiss (1979). Twenty-four reliability estimates were computed, one for each worker characteristic rated. The ratings were reliable to the extent that the within-task category ratings variance was significantly less than the between-task categories variance. The specific model used to estimate the reliability of the ratings treated raters as a fixed effect (i.e., did not seek to generalize the results beyond those particular raters) and used the mean rating as the unit of analysis. The analysis sought only to estimate the reliability of the ratings as they were used in the study. Because the average rating would be used to link worker characteristics to task categories, the reliability of the average rating, rather than of the raters <u>per se</u>, was estimated.

The reliability of the ratings for each worker characteristic is shown in Table 3.4. Reliabilities range from .97 (Written Communication, Physical Ability) to .60 (Tolerance of Negative Job Aspects) with an average of .88, indicating that the ratings were very reliable and therefore it is unlikely that the ratings would have differed appreciably should the supervisors have made the ratings a second time.

Table 3.2

31 Task Categories for Corrections Officers

Booking, Receiving, and Releasing

1. Obtaining information for booking purposes and filling out booking forms.
2. Searching, photographing, fingerprinting prisoners.
3. Logging inmate property, money; issuing IDs, clothing, and housing.
4. Reviewing court orders, dispositions; computing "good time", "work time", and release dates.
5. Performing paperwork for releases.

Escorting, Transporting

6. Verifying inmate identity.
7. Applying appropriate restraints.
8. Escorting inmates to desired locations inside or outside facility.

Record Keeping, Report Writing

9. Keeping routine records of inmate movement, personal data, activities, and authorized visitors.
10. Writing reports dealing with inmate behavior and officer response (e.g., incident and disciplinary reports).

Supervising Non-inmate Movement, Visitors

11. Admitting, releasing, and logging inmates' visitors.
12. Screening, searching visitors, and supervising visits.
13. Searching articles left by visitors for inmates.

Searching and Securing

14. Watching for potential disturbances, reporting emergencies, dispatching help, investigating disturbances.
15. Locking and unlocking cells and gates; taking and verifying inmate counts and location and status of inmates; counting/logging keys.
16. Searching inmates; shaking down cells and dorms; performing security checks/rounds.

Table 3.2 (continued)

Supervising Inmates

17. Supervising inmate meals and recreation.
18. Supervising phone calls, inspecting mail.
19. Supervising inmate work, inspecting cells for cleanliness.
20. Getting inmates up and ready for work, court, etc.
21. Passing out supplies, bedding, medication.

Communicating

22. Exchanging information with other shifts, control, maintenance crew, inmates, and visitors; making announcements, answering questions, conducting tours.

Physically Demanding Tasks

23. Physically subduing violent inmates; separating fighting inmates.
24. Lifting or dragging heavy objects (inmates or equipment).
25. Walking or standing extensively, running to scene of disturbance.

Miscellaneous Tasks

26. Administering CPR and first aid; extinguishing fires.
27. Dealing with emergencies: evacuating inmates or searching for missing/escaped inmates, on-the-spot counseling.
28. Attending staff and training meetings.
29. Reading logs, reports, memos, court documents.

Supervising Other Detention Facility Personnel

30. Assigning and evaluating work of other personnel.
31. Training or assisting other personnel.

Table 3.3 Average Importance Ratings of 24 Worker Characteristics for 31 Task Categories

Rating Scale:

1 = NOT IMPORTANT
2 = LOW IMPORTANCE
3 = MEDIUM IMPORTANCE
4 = HIGH IMPORTANCE
5 = VERY HIGH IMPORTANCE

JOB TASK CATEGORIES	Task No.	Facility-Specific Knowledge	Knowledge of the Legal System	Psychologist Training	Street Knowledge	First Aid	Using Firearms	Operating Communications Equipment	Oral Communication	Reading	Written Communications	Quantitative Skill	Observational Acuity	Driving	Self-Defense/Physical Restraint of Others	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Flexibility	Tolerance of Negative Aspects of the Job	Searching	Vision, Hearing Abilities	Physical Abilities
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
I.	1	3.7	2.9	2.8	2.9	2.6	1.3	2.9	4.5	4.1	4.2	2.8	3.9	1.2	2.7	3.6	3.5	3.7	3.3	3.1	3.4	3.9	3.0	3.8	3.0
Booking, Receiving, and Releasing	2	3.9	2.5	2.9	2.8	2.1	1.2	2.2	3.8	3.1	2.9	2.0	4.1	1.1	3.5	3.5	3.6	3.7	3.5	3.0	3.2	3.8	4.9	3.8	3.5
	3	4.3	3.2	2.1	2.4	2.0	1.2	2.0	3.4	3.4	3.9	3.9	3.8	1.0	2.7	2.8	3.4	3.0	3.1	3.1	3.1	3.6	3.2	3.4	3.0
	4	3.9	4.4	1.7	1.4	1.1	1.0	1.6	2.8	4.8	4.4	4.6	2.8	1.0	1.0	1.4	3.4	3.0	2.8	3.2	2.6	3.0	1.0	3.4	1.9
	5	4.2	4.0	1.7	1.3	1.2	1.0	2.0	2.5	4.4	4.5	3.5	2.9	1.0	1.0	1.8	3.4	2.9	2.5	2.9	2.3	3.1	1.2	3.1	2.0
II.	6	4.0	2.6	2.5	2.6	1.1	1.0	2.6	4.1	3.9	3.2	1.7	4.1	1.1	1.9	2.8	3.4	2.8	2.8	3.0	2.9	3.1	2.3	3.8	2.2
Escorting, Transporting	7	4.3	3.3	3.3	3.4	2.1	1.3	2.2	3.5	2.1	2.0	1.4	4.0	1.2	4.5	3.4	3.6	3.8	4.1	3.5	3.4	3.7	3.7	3.8	4.3
	8	4.5	2.8	3.3	3.4	1.9	2.7	3.3	3.9	2.2	2.0	1.9	4.5	3.4	3.8	3.6	3.5	4.0	4.0	3.1	3.7	3.6	4.0	4.5	4.5
III.	9	4.1	3.6	2.2	2.1	1.1	1.0	2.6	3.0	4.1	4.4	3.0	3.6	1.2	1.6	2.2	3.5	3.1	2.6	3.1	3.0	3.2	1.7	3.3	2.1
Record Keeping, Report Writing	10	4.2	3.9	2.9	2.7	1.1	1.0	1.9	3.1	4.0	4.8	2.3	4.3	1.0	2.0	3.2	3.6	3.2	3.3	3.2	3.5	3.3	1.7	3.2	2.0
IV.	11	4.2	3.6	3.0	3.0	1.4	1.0	2.4	4.0	3.7	3.7	2.2	4.0	1.0	2.4	3.7	3.4	3.3	3.3	3.1	3.4	3.2	2.9	4.0	2.6
Supervising Non-inmate Movement Visitors	12	4.4	3.8	3.6	3.4	1.5	1.1	2.5	4.0	3.1	3.0	1.9	4.2	1.0	2.8	4.0	3.5	3.8	3.7	3.3	3.5	3.6	4.5	4.1	3.6
	13	4.0	3.4	1.9	3.1	1.0	1.0	1.5	2.0	2.3	2.1	1.9	4.3	1.0	1.4	2.0	3.6	2.6	2.5	2.8	2.5	3.2	4.8	4.0	2.1
V.	14	4.4	3.3	4.0	3.4	2.4	1.3	4.0	4.3	2.5	3.0	1.6	4.7	1.2	3.5	4.0	4.0	4.2	4.1	3.6	3.6	3.8	3.5	4.5	4.3
Searching and Securing	15	4.6	2.8	2.8	2.4	1.5	1.0	3.1	3.4	3.1	3.3	4.2	4.5	1.0	2.6	3.0	3.8	3.9	3.8	3.3	3.0	3.7	3.0	4.2	3.3
	16	4.6	3.5	3.5	3.2	1.7	1.3	2.7	3.2	2.5	2.6	3.2	4.8	1.3	3.3	3.6	4.0	4.0	3.8	3.3	3.5	4.0	4.8	4.5	4.0
VI.	17	4.2	2.3	3.5	2.7	2.5	1.1	2.9	3.8	1.7	1.9	3.0	4.3	1.1	3.2	3.6	3.7	3.8	3.9	3.3	3.3	3.6	2.9	4.2	3.4
Supervising Inmates	18	3.6	3.9	2.6	3.2	1.5	1.0	2.0	2.5	4.0	2.6	2.0	4.3	1.0	1.7	2.8	3.4	3.1	3.3	3.1	2.9	3.4	3.2	3.8	2.0
	19	4.1	2.5	3.5	2.8	2.1	1.0	2.3	4.5	2.2	1.9	1.8	4.4	1.4	2.8	4.0	3.7	3.8	4.0	3.7	3.8	3.9	3.8	4.0	3.3
	20	3.9	2.7	3.3	2.1	1.6	1.0	2.2	4.3	2.2	1.7	1.9	4.0	1.5	2.8	3.7	3.7	3.6	3.7	3.1	3.4	3.7	2.9	3.8	3.0
	21	4.2	3.3	2.9	2.3	2.3	1.0	2.0	3.9	3.3	2.3	3.4	3.9	1.0	2.1	3.3	3.5	3.6	3.7	3.3	3.5	3.8	2.8	3.8	2.7
VII. Communicating	22	4.7	3.8	3.6	3.3	1.7	1.0	3.6	4.7	4.0	3.8	2.6	3.8	1.5	1.9	4.3	4.1	3.8	4.0	3.8	4.1	3.7	2.0	4.0	2.9
VIII.	23	4.0	3.7	3.8	3.3	2.7	1.5	2.3	4.3	1.5	1.9	1.5	4.3	1.1	4.8	3.8	4.0	4.5	4.4	3.3	3.8	4.3	3.0	4.4	4.8
Physically Demanding Tasks	24	2.7	1.8	2.0	1.5	2.3	1.0	1.5	2.1	1.3	1.2	1.4	3.0	1.1	2.6	2.0	3.0	3.1	2.7	2.6	2.9	3.4	1.7	2.8	4.9
	25	3.1	1.6	2.3	2.0	1.8	1.0	2.1	2.4	1.5	1.5	1.2	3.5	1.1	3.2	2.1	3.9	4.4	3.5	3.1	3.5	3.9	2.1	3.9	4.8
IX.	26	4.3	3.1	2.7	1.7	4.7	1.0	3.0	4.0	2.2	2.0	1.7	4.2	1.3	2.3	2.7	4.1	4.2	3.8	3.6	3.5	3.9	1.9	4.3	4.7
Miscellaneous Tasks	27	4.7	3.5	4.0	3.4	3.1	1.8	4.0	4.5	2.8	2.6	3.2	4.5	1.6	3.8	4.0	4.2	4.4	4.4	4.0	4.0	3.9	4.5	4.4	4.6
	28	3.3	2.9	2.9	1.8	1.4	1.4	1.7	4.2	3.5	3.6	2.3	3.4	1.1	1.5	3.7	3.9	3.3	3.4	3.4	4.0	3.3	1.2	3.4	2.0
	29	3.6	3.6	2.3	2.0	1.2	1.1	1.6	2.7	4.7	4.2	2.7	3.7	1.0	1.0	2.8	3.8	3.0	3.1	3.4	3.2	3.3	1.1	3.7	1.7
X.	30	4.2	3.2	4.0	2.0	1.3	1.4	2.2	4.5	4.3	4.4	2.7	4.2	1.1	1.3	4.1	4.3	3.9	4.0	3.2	4.1	3.4	1.6	3.3	2.4
Supervising Other Detention Facility Personnel	31	4.2	3.9	3.9	2.5	3.1	2.3	3.3	4.6	4.1	4.5	3.1	4.2	1.9	3.2	4.3	4.1	3.8	4.2	3.8	4.1	3.7	2.7	3.7	3.2

Table 3.4

Reliabilities and Ratings Summary
for Corrections Officers

Worker Characteristic	Rating rel.	Strength of Task Relationships[b] Low	Mod.	High
A. Facility-specific knowledge	.83	0	3	28
B. Knowledge of the legal system	.91	0	19	12
C. Psychological training[a]	.92	0	23	8
D. Street knowledge	.89	3	28	0
E. First aid	.93	13	17	1
F. Using firearms	.88	28	3	0
G. Operating communications equipment	.90	2	26	3
H. Oral communication[a]	.94	0	12	19
I. Reading[a]	.96	3	15	13
J. Written communication[a]	.97	2	17	12
K. Quantitative skill[a]	.94	4	23	4
L. Observational acuity[a]	.85	0	5	26
M. Driving[a]	.91	28	3	0
N. Self-defense/restraint of others	.95	6	21	4
O. Social skills[a]	.93	1	15	15
P. Motivation/dependability[a]	.72	0	8	23
Q. Stress tolerance[a]	.87	0	13	18
R. Dominance/self-esteem[a]	.89	0	14	17
S. Non-autonomy[a]	.61	0	25	6
T. Flexibility[a]	.84	0	19	12
U. Tolerance of negative job aspects[a]	.60	0	12	19
V. Searching	.96	4	19	8
W. Vision and hearing[a]	.85	0	9	22
X. Physical ability[a]	.97	0	21	10

[a]Before-hire worker characteristic

[b]See text (p. 18) for full descriptions

Setting Linkage Criteria

Once the task category to worker characteristic ratings were collected and determined to have been reliable, it was necessary to document the results and to define the level of importance (i.e., average rating) at which a worker characteristic would be considered linked to a task category. For purposes of summarizing the strength of the task category to worker characteristic linkages, the 5-point rating scale was trichotomized into low (average rating of 1.0 - 1.5), moderate (1.5-3.5). and high (3.5 - 5.0). Table 3.4 summarizes the number of task categories linked to each worker characteristic at each of the three linkage strengths. As can be seen, for approximately 3/4 Of the worker characteristics the task category linkages were of a single strength. For the remaining 1/4, the linkages were approximately evenly divided between two strengths.

It was desired to set the criterion for linking a worker characteristic to a task category as high as possible so that the tests developed to measure each linked worker characteristic would provide accurate predictions of task performance. Developing a test to measure a worker characteristic that is only weakly linked to the task category the test is to predict, would not result in a selection battery likely to provide accurate predictions of job performance. Initially, the criterion was set at 4.0, that is, only those worker characteristics with average importance ratings of 4.0 or higher would be considered linked to a task category. As will be discussed however, it was necessary to relax this criterion to ensure that all of the tasks were linked to at least one predictor. At the time these data were being collected and analyzed, the core job had not been defined (see Job Analysis Report for details). As a result, the relative importance of the 194 job tasks was not known. Rather than risk overlooking a task category that might later have been found to be an important, or even a critical aspect of the corrections officer job, the linking criterion was adjusted so that every task category, therefore every task, would be linked to at least one worker characteristic for which a test could be developed.

Linking Tests to Worker Characteristics

In much the same way that worker characteristics were linked to task categories, it was necessary to link the tests being considered for inclusion in the experimental selection battery to the same set of worker characteristics (recall that it is through the worker characteristics that tests were linked to the tasks they were to predict). The first step in this process was to compile a list of tests having potential to predict the job performance of corrections officers. Based upon research reports located during the literature review, the job analysis materials and results, and the experience of PDI staff members, a list of 72 tests was compiled. For each test, a one-page description was constructed describing the characteristic it measured, its reliability and, if available, a summary of its validity for similar jobs.

Eleven psychologists from PDI and its sister organization Personnel Decisions Research Institute then rated the degree to which each of the 72 tests would be a useful measure of each of the corrections officer worker characteristics. The full set of 24 worker characteristics was narrowed to 16 by deleting those characteristics which, on the basis of the job analysis data, were found as not being necessary before hire. This was done because any worker characteristic not necessary at the time of hire (i.e., that could be trained on the job) would not be an appropriate selection standard and, therefore, linking tests to those worker characteristics would have served no useful purpose. In order that no worker characteristic be prematurely rejected from consideration, this narrowing of the list of worker characteristics was done as conservatively as possible. Before-hire worker characteristics were defined as those having at least one constituent KSA which, according to at least 40% of corrections officer supervisors, should be developed before hire (i.e., at least 40% of supervisors rated the KSA 0 or 1 on the 0 to 4 "when developed" JAQ scale).

Ratings of the usefulness of each test as a measure of each worker characteristic were made on an 8-point scale ranging from "not at all useful" (rating of 0), through "almost useless" (ratings of 1 or 2), "of some use" (ratings of 3, 4, or 5), "highly useful" (ratings of 6 or 7), to the worker characteristic being "entirely measured" by the test (rating of 8). Table 3.5 shows the average usefulness of each test as a measure of each of the 16 before-hire worker characteristics.

The reliability of the test to worker characteristic ratings was determined using the same methods described for determining the reliability of the task category to worker characteristic ratings. The reliability analyses indicated that, with the exception of the Physical Ability worker characteristic, the ratings were all quite reliable (range = .72 to .97, average = .86). The reliability of the Physical Ability ratings (intraclass r = .08) at first glance appears quite dismal, but this apparent unreliability is a statistical artifact rather than an indication of unreliable ratings. The intraclass correlation estimate of reliability is an index of the relative amount of variation in the ratings data. As the data in Table 3.5 reveal, there was almost no variation in the psychologists' ratings of the usefulness of the 72 tests as measures of the Physical Ability worker characteristic (i.e., no tests were rated as useful measures of physical abilities). Because there was essentially no variation in the ratings, the intraclass correlation must be near zero: Nevertheless, the ratings were made reliably in the sense that all raters agreed as to the usefulness (lack of usefulness, actually) of the 72 tests as measures of corrections officer physical abilities.

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 16 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
1. EAS Ver. Comp.	3.3	5.0	5.0	5.4	1.4	1.3	.1	1.4	.4	.4	1.2	.1	.4	1.4	.1	0
2. EAS Num. Ab.	1.5	.4	1.3	1.2	7.5	.3	.5	.1	.2	.1	.4	.2	.2	.5	.1	0
3. EAS Visual S & A	.3	.1	1.4	1.0	1.2	2.3	1.3	.1	.2	.1	.2	.4	.9	.1	1.4	0
4. EAS Space Vis.	.1	.2	.4	.3	1.1	1.0	2.4	.1	.1	.1	.1	.2	.2	.1	1.4	.2
5. EAS Ver. Reas.	3.4	3.3	3.8	3.6	.9	1.1	.3	1.1	.1	.2	.6	.4	.1	1.0	.1	.1
6. EAS Symb. Reas.	2.1	1.1	1.5	1.8	2.0	.6	.3	.3	.1	.1	.4	.3	.1	1.1	.1	.1
7. DAT Abst. Reas.	2.7	1.4	2.0	1.6	2.3	1.2	.4	.4	.1	.1	.1	.3	.1	1.1	.4	.1
8. DAT Lang. Usage	2.1	4.2	4.1	6.3	1.0	.3	.1	.7	.1	0	.4	.2	.1	.3	.1	.1
9. DAT Ver. Reas.	3.8	3.6	3.6	3.8	.9	1.1	.1	.7	.1	.1	.4	.3	.1	.9	.1	.1
10. EEI Follow D.	.8	1.8	1.7	.8	1.0	.7	1.5	.2	.4	.3	.2	2.2	.9	.3	.9	.1
11. EEI Tables & Gra.	.8	.5	1.4	.6	1.3	.9	.4	.1	.1	.1	.2	.9	.4	.2	1.3	.1
12. ETS Hid. Fig.	.2	.1	.4	.4	.3	2.3	1.3	.1	.1	.1	.2	.2	.2	1.6	1.6	.1
13. ETS Card Rotat.	.2	.1	.5	.3	.6	1.9	1.3	.1	.1	.1	.1	.1	.1	1.1	1.1	.1
14. ETS Map Mem.	.4	.2	.6	.3	.5	3.1	1.7	.1	.1	.1	.1	.2	.1	.2	1.2	.1

[a] Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 16 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
15. ETS Bldg. Mem.	.3	.1	.4	.2	.4	3.0	1.4	.1	.1	.1	.1	.1	.1	.2	1.2	.1
16. ETS Follow Dir.	.8	2.0	2.2	.9	.7	.7	.8	.1	.1	.1	.1	1.6	.8	.6	.3	.1
17. Watson-Glaser	3.4	2.9	4.1	3.2	1.4	1.5	.4	1.2	.3	.1	.6	.4	.1	.9	.1	.1
18. CO Reading	2.0	2.3	7.0	2.8	1.0	.6	.4	.3	.1	.1	.1	.3	.1	.3	.2	.1
19. Otis-Lennon	3.2	2.8	3.4	3.0	3.4	1.5	.4	1.0	.3	.1	.4	.3	.1	1.0	.2	.1
20. Wesman Verb.	3.2	3.0	3.9	3.4	1.8	1.4	.3	1.1	.3	.1	.2	.3	.1	.9	.1	.1
21. CAB Ma	1.3	.9	1.4	1.3	.8	2.0	1.0	.3	.1	.1	.1	.6	.4	.2	.4	.1
22. CPI Do	.8	2.3	.3	.9	.4	1.1	.4	3.2	1.2	1.6	6.6	1.6	1.0	1.4	.1	.1
23. CPI Sy	1.3	2.4	.5	.3	.2	1.4	.1	5.3	.8	.8	2.6	1.1	.5	1.4	.1	.1
24. CPI Wb	1.2	1.0	.3	.1	.1	.7	.1	3.4	2.8	2.9	3.5	1.9	3.0	1.4	.1	.1
25. CPI Re	.9	.4	.3	.4	.1	.7	.7	1.7	5.9	1.9	1.4	4.4	2.7	1.0	.1	.1
26. CPI So	.5	.4	.3	.2	.1	1.4	.9	2.4	5.3	2.0	1.4	4.0	3.2	1.5	.1	.1
27. CPI Sc	1.3	1.0	.3	.2	.1	.5	1.1	2.4	3.2	4.3	1.6	3.0	2.3	1.6	.1	.1
28. CPI To	1.7	1.5	.4	.3	.3	1.4	.5	4.4	1.7	2.2	1.7	3.3	2.2	2.8	.1	.1

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 16 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
29. CPI Ac	1.8	1.4	1.3	1.3	.8	.9	.4	1.4	3.4	1.4	1.0	3.9	2.4	2.3	.1	.1
30. CPI Fx	1.0	1.2	.4	.4	.2	.4	.5	1.3	1.0	1.4	.7	2.8	2.0	4.4	.2	.1
31. CPI Ie	3.0	2.4	2.4	2.3	1.9	1.6	.6	1.4	1.7	1.3	1.3	1.4	.7	1.7	.1	.1
32. CPI Ai	1.8	1.7	1.5	1.6	1.3	.9	.4	1.4	1.7	.8	1.2	1.7	.8	2.0	.1	.1
33. MMPI L	.8	.6	.2	.1	.1	.6	.1	1.4	.8	1.4	1.6	1.2	.7	.7	.1	.1
34. MMPI F	.6	.8	.8	.3	.2	.7	.1	1.5	1.6	2.5	.8	.6	.7	.1	.1	.1
35. MMPI K	.2	.5	.2	.1	.1	.8	.1	1.9	1.6	1.6	1.1	1.0	.8	.9	.1	.4
36. MMPI 1	.8	.4	.2	.1	.1	1.4	.3	2.2	1.7	2.4	1.4	.2	1.0	.3	.2	.1
37. MMPI 2	.4	.7	.2	.1	.1	.7	.5	3.0	1.5	2.6	2.3	0	1.1	.9	0	.1
38. MMPI 3	.4	0	.1	0	.2	.4	.3	1.9	1.8	3.5	.9	.6	2.0	.4	0	0
39. MMPI 4	.3	.8	.1	0	0	.4	1.2	2.7	3.3	1.3	1.6	2.8	1.5	.3	0	0
40. MMPI 5	.6	.3	.3	.2	.2	1.3	.2	1.5	.8	.5	1.4	.7	.4	.7	0	0
41. MMPI 6	.6	.2	.1	0	0	.8	.2	1.1	.3	.9	.9	.1	.7	0	0	0
42. MMPI 7	.8	1.3	.2	.7	.1	1.5	0	1.7	2.1	2.7	2.8	2.0	2.0	1.8	0	0

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 16 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
43. MMPI 8	.6	.9	.1	0	0	.9	.2	2.4	1.5	2.0	1.8	1.7	1.9	.2	0	0
44. MMPI 9	.5	.7	.4	.3	.3	.7	.8	2.0	2.1	2.2	2.2	.6	.7	.7	0	0
45. MMPI 0	.2	1.9	.1	.1	.1	1.1	0	4.6	1.4	1.7	3.0	.2	.2	1.1	0	0
46. MMPI 0-H	.7	.2	0	0	0	.2	.7	2.1	.9	2.8	.9	2.3	1.3	1.2	0	0
47. GPI Coop.	.4	.7	.1	.1	.1	.8	.2	4.4	1.5	2.2	1.4	1.2	1.1	.9	0	0
48. GPI Conscient	.1	0	0	.4	.2	.6	1.4	.4	2.3	1.4	1.2	4.4	1.0	1.4	0	0
49. GZTS Gen. Act	.4	.8	.2	.3	.2	0	.4	1.6	1.5	.7	1.9	.9	1.0	.3	0	.2
50. GZTS Ascend.	.4	2.4	.2	.2	.2	.4	0	3.5	1.4	1.4	5.4	1.6	.5	.2	0	0
51. GZTS Emot. Stab	.6	1.4	.1	.1	.1	.5	.4	2.4	2.1	3.9	1.6	.7	2.3	1.4	0	0
52. GZTS Objectiv.	.6	.9	.2	.2	.2	.8	.2	2.3	1.3	4.1	2.4	1.1	2.2	1.6	0	0
53. GZTS Pers. Rel.	.4	.4	.1	.1	.1	.4	.2	3.6	1.5	1.2	1.4	1.6	1.5	1.2	0	0
54. GZTS Restraint	.4	.5	.1	.2	.2	.7	1.4	1.1	3.5	2.2	1.0	1.6	1.3	.6	0	0
55. DPQ Unfrien	.1	.1	0	0	0	.8	.4	3.2	1.4	2.0	1.0	1.4	1.8	1.0	0	0
56. DPQ Authorit.	.4	.1	0	0	0	.6	.4	1.4	2.9	1.1	.7	3.7	2.3	1.4	0	0

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 16 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
57. DPQ Stress R.	.6	.8	.1	.1	.1	.4	.8	1.4	1.8	5.3	1.9	.7	2.8	1.6	0	0
58. DPQ Aggression	.2	.2	0	0	.5	.2	1.6	3.4	2.2	2.3	1.6	1.3	1.3	1.4	0	0
59. DPQ Impulsive	.5	.8	0	0	0	.2	.7	1.7	3.0	1.5	.5	2.1	1.6	2.2	0	0
60. DPQ Control	.3	.1	0	0	.2	.2	.4	1.1	3.2	1.5	.4	2.8	1.5	1.6	0	.2
61. DPQ Tradition	.3	.1	0	0	0	.2	.4	1.2	3.6	.4	1.3	3.2	3.0	2.2	0	0
62. APQ	.1	1.2	0	0	0	.4	.2	2.9	.3	4.9	1.4	.4	2.7	1.1	0	0
63. BQ Soc. Inr	.4	1.8	0	0	0	.4	0	4.0	.2	1.2	2.8	.4	.4	.5	0	0
64. BQ Agg/Indep	.1	1.7	0	.2	0	.4	0	2.3	.3	.5	4.0	1.7	1.1	1.0	0	0
65. BQ Soc. Des.	.2	.4	0	.2	0	.4	.4	1.6	1.0	1.3	2.3	1.7	1.7	.3	0	0
66. BQ Soc. Lead.	.1	1.8	.2	.1	.2	.3	0	4.4	.9	1.2	3.9	.8	.5	.2	0	0
67. BQ Adjust.	.5	.6	0	0	.2	.4	.3	2.5	1.2	2.5	2.9	.6	1.2	.4	0	.1

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 16 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
68. BQ Exp. Neg.	.4	.5	0	.1	0	.1	0	2.4	.5	1.4	.8	1.2	1.2	.1	0	0
69. Self-Esteem	.6	1.0	.3	.3	.3	.2	.2	1.9	.7	1.8	3.5	.2	.9	.4	0	.2
70. Empathy	1.3	.8	0	0	0	2.4	.1	4.5	.5	.5	1.5	.8	.9	1.9	0	0
71. CO Interest	.7	.5	0	0	0	1.3	0	2.3	1.8	2.1	1.1	1.9	1.9	1.5	0	0
72. Occup. Role	.8	.7	.1	0	.1	.1	0	1.7	1.2	.7	.7	2.2	1.7	.6	0	.3

Table 3.5 (continued)

Employee Aptitude Survey

1. Verbal comprehension
2. Numerical ability
3. Visual speed and accuracy
4. Space visualization
5. Verbal reasoning
6. Symbolic Reasoning

Differential Aptitude

7. Abstract reasoning
8. Language usage
9. Verbal reasoning (analogies)

Edison Electric Institute

10. Following directions
11. Tables and graphs

Educational Testing Service

12. Hidden figures
13. Card rotations
14. Map memory
15. Building memory
16. Following directions

17. Watson-Glaser Critical Thinking Appraisal

18. Correctional Officer Reading Comprehension Test

19. Otis-Lennon Mental Ability Test

20. Wesman Personnel Classification Test, Part 1

21. Comprehensive Ability Battery Rote Memory

Table 3.5 (continued)

California Psychological Inventory

22. Dominance
23. Sociability
24. Well-being
25. Responsibility
26. Socialization
27. Self control
28. Tolerance
29. Achievement via conformance
30. Flexibility
31. Intellectual efficiency
32. Achievement via independence

Minnesota Multiphasic Personality Inventory

33. Lie scale
34. Infrequency
35. Validity
36. Social introversion
37. Hypomania
38. Schizophrenia
39. Psychasthenia
40. Paranoia
41. Masculinity-femininity
42. Psychopathic deviate
43. Conversion hysteria
44. Depression
45. Scale 0
46. Scale 0-H

Gordon Personal Inventory

47. Cooperativeness
48. Conscientiousness

Guildford-Zimmerman Temperament Survey

49. General activity
50. Ascendancy
51. Emotional stability
52. Objectivity
53. Personal relations
54. Restraint

Table 3.5 (continued)

Multidimensional Personality Questionnaire

55. Unfriendly world
56. Authoritarianism
57. Stress Reaction
58. Aggression
59. Impulsiveness
60. Control
61. Traditionalism

62. **Activity Preference Questionnaire**

Owens Biographical Questionnaire (unpublished)

63. Social introversion
64. Aggressiveness/independence
65. Social desirability
66. Social leadership
67. Adjustment
68. Expressing negative emotion

69. **PDRI Self Esteem**

70. Hogan's Empathy

71. Correctional Officers' Interest Blank

72. PDRI Occupational Role Preferences.

To summarize the results of the test to worker characteristic ratings, the 8-point scale was reduced to four categories. Tests whose average usefulness rating fell between 0.0 and 0.5 were said to have no relationship to the worker characteristic, tests with ratings between 0.5 and 2.5 were said to have a low relationship, between 2.5 and 5.5 a moderate relationship, and tests with an average rating of 5.5 or above were said to be highly related to the worker characteristic. Table 3.6 presents the number of tests having each degree of relationship to each of the 16 before-hire worker characteristics. As can be seen, most worker characteristics were judged moderately related to several tests. There were two exceptions to this finding: Driving and Physical Abilities, both of which were essentially unrelated to any of the 72 tests.

Linking Tests to Task Categories

Once the task category to worker characteristic and worker characteristic to test ratings were made, the stage was set to link tests to task categories through the intervening worker characteristics. This linking of tests to task categories required the simultaneous examination of both sets of ratings. The strategy adopted for linking tests to tasks was to set criteria that were as stringent as possible, while at the same time ensuring that at least one test was linked to each task category. Procedurally, each task category was considered in turn and all of the worker characteristics whose average ratings met or exceeded the specified criterion (see Table 3.7) were linked to that task category. Each of the worker characteristics so linked were themselves examined and all of the tests whose average usefulness rating met or exceeded the second criterion were linked to the task. For example, for the first task category (Obtaining information for booking purposes and filling out booking forms), three worker characteristics (Oral Communication, Reading, and Written Communication) had average importance ratings of 4.0 or higher. For Oral Communications, one test (EAS Verbal Comprehension) received an average usefulness rating of 5.0 or higher, for Reading, two tests (EAS Verbal Comprehension and CO Reading) were rated 5.0 or higher, and for Written Communications, two tests (EAS Verbal Comprehension and DAT Language Usage) were rated 5.0 or higher. For this example, eliminating duplicates yielded a set of three tests to be used in the experimental selection battery.

Table 3.6

Reliabilities and Ratings Summary of
Test to Worker Characteristic Ratings

Before-hire Worker Characteristic	Strength of Task Relationships[a] None	Low	Med.	High	Rating rel.
A. Knowledge of psychology	28	36	8	0	.88
B. Oral communication	22	43	7	0	.89
C. Reading	49	15	7	1	.97
D. Written communication	51	13	7	1	.96
E. Quantitative skill	49	21	1	1	.95
F. Skill in observing behavior	23	47	2	0	.72
G. Driving	47	25	0	0	.74
H. Social skills	15	40	17	0	.91
I. Motivation/dependability	25	36	10	1	.90
J. Stress tolerance	22	38	12	0	.89
K. Dominance/self-esteem	19	42	10	1	.90
L. Non-autonomy	24	38	10	0	.86
M. Tolerance of neg. job aspects	21	45	6	0	.78
N. Flexibility	22	48	2	0	.74
O. Vision and hearing	64	8	0	0	.82
P. Physical abilities	72	0	0	0	.08

[a]See text (p. 29) for full descriptions.

Table 3.7

Linkage Criteria and Summary for Test to Worker Characteristic and Worker Characteristic to Task Category Ratings

Task Category[a]	Task Category to Worker Characteristic		Worker Characteristic To Test	
	Criterion	N Linked	Criterion	N Linked
1	4.0	3	5.0	5
2	4.0	1	3.0	2
3	3.8	3	5.0	3
4	4.0	3	5.0	5
5	4.0	2	5.0	4
6	4.0	2	5.0	1
7	4.0	3	5.0	2
8	4.0	5	5.0	3
9	4.0	2	5.0	4
10	4.0	3	5.0	4
11	4.0	3	5.0	1
12	4.0	4	5.0	2
13	3.5	3	5.0	2
14	4.0	9	5.0	7
15	4.0	3	5.0	7
16	4.0	6	5.0	3
17	3.5	9	5.0	7
18	4.0	2	5.0	2
19	4.0	5	5.0	4
20	4.0	2	5.0	1
21	3.8	4	5.0	1
22	4.0	7	5.0	8
23	4.0	8	5.0	6
24	3.5	1	5.0	0[b]
25	4.0	2	5.0	1
26	4.0	6	5.0	4
27	4.0	11	5.0	7
28	4.0	2	5.0	1
29	4.0	2	5.0	4
30	4.0	9	5.0	10
31	4.0	8	5.0	10

[a]See Table 3.2 for Task Category Labels.

[b]Task Category 24 was strongly linked to only one worker characteristic (Physical Ability) which was not linked to any of the 72 tests.

Initially it was hoped that two criteria could be set, one used to link all task categories to worker characteristics, and the other used to link all tests to worker characteristics. As Table 3.7 indicates, in a few instances it was necessary to decrease the magnitude of one or the other criteria in order that at least one test be linked to each task category. Once all test to task category linkages were made, the list of linked tests was revised to eliminate duplicates (due to worker characteristics being linked to more than one task), yielding the final set of 10 tests (see Table 3.8) that comprise the experimental selection battery for corrections officers.

Predictor Development

As was detailed above, PDI and PDRI staff psychologists rated the usefulness of 72 tests of cognitive or noncognitive abilities as measures of the worker characteristics necessary to successfully perform the corrections officer job. Analysis of the ratings showed them to be very reliable and indicated that 10 of the 72 tests (See Table 3.8) were likely to be useful predictors of job performance. Four of the tests were measures of noncognitive abilities. The remaining six were measures of cognitive abilities. The contract between STC and PDI allowed for the use of existing noncognitive tests in the final selection battery, but required all cognitive ability tests to be newly developed so that the copyright to the tests could be assigned to the California Board of Corrections. In accordance with the contract, the noncognitive measures were permissible for use in the selection battery "as is", contingent only upon appropriate permissions and royalty agreements being negotiated with the test publishers. The remaining tests, being measures of cognitive ability, could not be used in the final battery and so required the development of new tests designed to measure the same abilities.

Development of the new cognitive ability tests involved three steps. First, items were drafted and preliminary test booklets assembled. Second, the newly constructed tests were pretested in a sample of college students (administered with the marker tests they were designed to replace). Third, data from the pretest were used to refine the tests into a form ready for use in the experimental selection battery. Each of the three steps is detailed below.

Test Construction. In order that the new tests parallel as closely as possible the format and content of the marker tests they were to replace, the first step in test construction was to outline a test development plan for each test. The development plans (see sample in Figure 3.2) described the ability to be measured, the number and type of items to be included, whether the test was intended to be a speed or power test, the proposed time limit for the test, and any other comments that might be useful in guiding construction of the test. Once the test development plans were completed, test construction proceeded similarly for each of the five tests. Only five tests were developed because two

Table 3.8

Cognitive and Noncognitive Tests for the Experimental Selection Battery

Marker	Test Description
Cognitive Ability	
EAS Verbal Comprehension	A 30-item vocabulary test with a five-minute time limit designed to measure the ability to use words in oral and written communications.
EAS Numerical Ability	A 75-item arithmetic test in three parts with a total test 10-minute time limit. The test measures a person's facility with the arithmetic operations of addition, subtraction, multiplication, and division using problems combining integers, decimals, fractions, and percentages.
DAT Language Usage	A 50-item test in which persons must identify which of four sections of a sentence, if any, contains an error in punctuation, grammar, or capitalization. The time limit for this test is 25 minutes.
ETS Map Memory and ETS Building Memory	Each is a 24-item test with 12-minute time limits (6 minutes for each of two parts) designed to measure one's "ability to remember the configuration, location, and orientation of figural material". The tests differ only in the nature of the material to be memorized (map sections versus building locations).

Table 3.8 continued

Marker	Test Description
CO Reading	A 50-item reading test designed to measure how well one can read and remember what was read. The test consists of short passages, each followed by several questions.
Noncognitive Tests	
CPI Dominance	A 46-item subscale of the California Psychological Inventory that identifies individuals who behave in a dominant, ascendant manner and who, in interpersonal situations, can take initiative and exercise leadership in a forceful, self-confident manner.
CPI Sociability	A 36-item subscale of the California Psychological Inventory that indicates a person's tendencies to be outgoing and participative and to seek out and enjoy social contact without being shallow or superficial.
CPI Responsibility	A 42-item subscale of the CPI measuring a person's tendency to believe that life is best governed by reason and to behave in a responsible, self-disciplined manner.
MPQ Stress Reaction	A 26-item subscale of the Multidimensional Personality Questionnaire that indicates persons who are able to put fears and worries out of their minds easily, who quickly get over upsetting experiences, and who are not troubled by emotional turmoil or feelings of guilt.

Figure 3.2

Test Development Plan for Numerical Ability Test

Ability Measured
Mathematical knowledge and computational ability.

This test is to measure knowledge of the four basic mathematical operations (addition, subtraction, multiplication, and division). Applicants will be required to apply this knowledge to simple problems involving whole numbers, fractions, decimals, and percentages.

Marker Test
Employee Aptitude Survey Test 1.

Test Format
Test should consist of 50 multiple choice items, each with four distractors. This is intended to be a speeded test with a time limit, excluding instructions, of five minutes.

Comments
This test is intended to roughly parallel the EAS marker but with fewer items and a slightly different item mix. Distractors should be devised to be the results that would obtain from common errors. For example, a problem such as 1/3 x 1/2 might have as distractors 1/5, 2/5, 2/6, and the correct answer, 1/6. In this way, the test will tap both the knowledge (what to do) and computational (actually doing it) aspects of numerical ability.

---

of the marker tests (ETS Map Memory, ETS Building Memory) were so similar in content that a single test was developed to replace both of these marker tests.

Using the test construction plans as blueprints, PDI staff members drafted 2 to 2 1/2 times the number of items needed for the final tests. The surplus of items ensured that the tests could be significantly revised and still contain sufficient items to be reliable and valid.

Once items were drafted, two preliminary forms of each test were assembled and scoring keys developed and checked. Item distractors were arranged so that the correct response occupied each possible position (i.e., A, B, C, D, or E) with approximately equal frequency. Each test was then reviewed by PDI staff members for obviously biased or unfair items and the scoring keys rechecked for accuracy. The resulting preliminary tests were printed in sufficient quantity for pretesting. Preliminary test construction yielded two forms of each of the five cognitive ability tests to be used in the experimental battery.

Pretesting the Preliminary Tests. Arrangements were made with instructors at four-year and community colleges to permit their students to serve as participants in the pretest sessions. In return the students received extra credit toward their class grades. Each preliminary test was administered, under the same conditions to prevail during actual test use, to groups of 15-30 students. No student completed more than two of the preliminary tests. Table 3.9 presents the number of students completing each of the preliminary tests. Two of the tests listed in Table 3.9 were not specifically targeted for the corrections officer position. The tests were developed as potential predictors of one of the two other jobs studied in this project. They are included in Table 3.9 because, although not targeted to the corrections officer job, they were part of the experimental selection battery administered to corrections officers. A single version of the experimental selection battery was prepared for use with all three jobs. This was considered necessary as validation testing for the three jobs occurred simultaneously, and it was critical to prevent any accidental cross administration of, say, the probation officer tests to corrections officers.

During pretesting, students were administered the appropriate marker test (under the time limits and instructions provided by the publisher), followed by both forms of the preliminary test. For the preliminary tests, students were instructed to mark the item they were working on at the end of the proposed time limit, but were then permitted to continue working until they had completed all items. This procedure allowed the time limits to be gauged as well as permitting the collection of data about every test item. Because the data were used to compute the item statistics that were in turn used to select

Table 3.9

Pretest Results and Item Statistics for the Experimental Selection Battery

Test	Number of Students Tested	Avg. Item Diff.	Item Test Corr.	N Items	rel.[b]	Validity[c]
Observational Memory	30	.90	.10	30	.44	.31
Numerical Ability	19	.74	.57	35	.92	.75
Vocabulary	21	.66	.37	35	.83	.82
Language Usage	29	.78	.51	35	.90	.88
Reading Comprehension	30	d	d	70	d	d
Analogies[a]	16	.78	.62	35	.93	.69
Verbal Reasoning[a]	25	.57	.26	35	.80	.72

[a]Not targeted for the corrections officer job -- see text (pp. 36) for details.

[b]Internal consistency via KR-20.

[c]Correlation with marker test.

[d]Data unavailable due to method of revision -- see text (pp. 40) for details.

the final items, it was important to ensure that psychometrically sound items were not eliminated from the final tests simply because they appeared near the end of a preliminary test. Such elimination might have occurred had the pretesting time limits been so short as to prevent students from attempting all of the items.

Item Analysis and Revision. Using the data collected in the pretest sessions, three statistics were computed for every preliminary test item. The difficulty of each item (proportion of persons passing the item) was computed as was the correlation of each item with the total score. Item-total correlation were computed as the point-biserial correlation between an item (scored dichotomously as pass or fail) and the score received on the sum of the remaining items. Finally, the frequency with which each distractor was chosen was determined.

The item statistics were used to select items to appear in the experimental selection battery. Prior to this step, the average number of items completed in the proposed time limit was computed for each test. These results showed that students were working the items quickly enough to permit a switch from speed to power measures of each of the cognitive abilities. Speed measures are those where the constituent items are so easy that, given sufficient time, everyone should be able to get every item correct. For speeded tests, time limits are set that are much shorter than is necessary for most people to complete all test items. Speed tests are predicated upon the idea that people with more or higher levels of the ability being measured will be able to work the items faster. Thus, a higher score on speed tests is an indication of better ability. On the other hand, power tests have time limits that permit most persons to complete all items. On power tests the items vary in difficulty with some being so easy that nearly everyone can get them correct while others are much more difficult.

To summarize, speed tests depend upon time limits to distinguish among persons of varying ability while power tests depend upon item difficulty. The advantage of power tests, especially in a project such as this, lies in their independence from time limits.

In order for any test to give an accurate measure of ability, or in this case the potential to be a successful corrections officer, the test must be administered under as nearly identical conditions as possible every time it is used. With speed tests, especially those with relatively short time limits, it is very difficult to ensure that the time limits are rigidly followed during every administration. Because they have no strict time limits, power tests are more desirable for projects such as this. As a result, each test in the experimental selection battery (with two exceptions discussed below) contained approximately the average number of items completed in the pretesting time limit. The total time allowed for the resulting battery was then

doubled. Thus, the final time limits were set so that virtually every person would be able to complete all items within the allotted testing time. In this manner, the tests were transformed from highly speeded to power tests, with an accompanying reduction in administrative complexity.

Once the desired number of items was determined for each test, the pretest data were used to screen items and to reject those that were not of appropriate psychometric quality. Several factors were considered in the course of this screening. For example, items for which the keyed alternative correlated negatively with total scores, or where the item-total correlation for the keyed alternative was less than that for a distractor were rejected. No initial screening was done on the basis of item difficulty. The items remaining after screening formed the final pool of psychometrically sound items from which item selection was made.

In order to ensure an overall difficulty level appropriate for the intended job applicant population, items were selected based upon their item difficulties. For example, five items with difficulties between 1.0 and .90, five with difficulties between .90 and .80, and so on, were chosen for each test. Within difficulty levels, items with the highest item-total correlations were selected.

Once the desired number of items had been selected, the reliability (KR-20) and validity (correlation with the published marker test) of each test were determined. If these were satisfactory, test construction was complete. If not, the correlation of each item with the marker test total score was computed (for all items in the final item pool), and this index used as the basis for item selection. That is, within a specific range of item difficulty, the items which correlated most highly with the marker total score were selected. Table 3.9 summarizes the psychometric properties of the tests that make up the experimental selection battery.

As was true for the preliminary tests, items in the final tests were arranged so that no patterns of correct answers were present and so that the correct alternative appeared in each possible position with approximately equal frequency. Finally, items were arranged (approximately) in ascending order of difficulty.

Concern that the tests might prove too difficult for the applicant population led to the inclusion of five additional items in each test. All of the additional items had item difficulties of 1.0 in the pretest sample. These items were included to ensure that each test had sufficient breadth to prevent restricting the range of test scores due to floor effects. "Floor effects" refers to a situation in which the items of a test are all too difficult for the test taking population. If this occurs, scores on the test will all be very low and there will

be a large number of persons who receive the same score. A large number of tied scores reduces the variance of the scores and puts a limit on the maximum possible validity of a test. In order not to limit the potential validity of the tests, several very easy items were added to each test.

Special Revision Methods. One of two tests not revised as described above was the Observational Memory Test. Rather than selecting a subset of the 30 items in the preliminary version of this test, it was decided to retain the test intact. Data on the pretest time limits showed that the test was easily completed in a minimal amount of time therefore little was to be gained by shortening the test. The time limit results, coupled with item statistics indicating that the items were too easy, led to a reduction in the suggested time limits. Rather than having two minutes to study the stimulus figures to be memorized, only one minute was allowed. The amount of time allowed to complete the ten items in each test section was also reduced from three to two minutes. In addition to the changed time limits, the items were made more difficult by increasing the complexity of the stimuli to be memorized and by requiring more subtle discriminations between correct and incorrect alternatives. No other revision of the test was undertaken and it remained the only test to depend upon speed rather than power to distinguish among candidates.

The other test for which a variation in the revision procedures was made was the Reading Comprehension test. Pretest data for the reading marker test (CO Reading) revealed that the test was too easy for the pretest sample, leading to very little variation in test scores. Much the same was true for two of the three parts of the preliminary reading test. As a result, the item statistics were judged too unreliable to serve as a guide to test revision. One section of the preliminary test (Part 2) in particular was clearly not working as intended. This part of the test required subjects to make inferences and draw conclusions from minimal amounts of information. The problem with this part of the test was that no clearly justifiable key could be agreed upon and subjects were about equally divided in their choice of distractors.

The experimental selection battery Reading Comprehension test was made up of the six reading passages that PDI staff judged to be most appropriate for the intended population. The 10 to 12 items following each passage were an approximately 75%-25% mix of "search and find" and reasoning items. The former required the recall of answers to factual questions, while the latter required simple inferences and conclusions to be drawn from the material presented in the passages. These reasoning items are the only remaining vestige of Part 2 of the preliminary scale. Enough items were included in each passage to allow the test to be further refined (based on item statistics and using the procedures described above) using data collected during the validation phase of the project.

Final Test Battery

The newly developed cognitive ability tests were assembled in three booklets. The first booklet contained the Observational Memory test. This test required strict adherence to time limits and so was kept separate from the other tests. The second booklet contained the Reading Comprehension test. A time limit of 30 minutes was set for this booklet. On the basis of the pretest data, 30 minutes was judged to be sufficient to allow virtually all incumbents to complete all test items. The test consisted of six passages, each with 10 to 12 items. Other than the 30-minute limit for the entire booklet, incumbents were not constrained in the amount of time they could spend on any one passage.

The third test booklet contained the Numerical Ability, Language Usage, Vocabulary, Analogies, and Verbal Reasoning tests. The Analogies and Verbal Reasoning tests, while not targeted specifically to the corrections officer job, were administered as part of the experimental selection battery. This was done for administrative convenience--permitting administration of the same test battery to the validation samples for all three jobs covered in the project. Because testing was to occur simultaneously for all jobs, it was feared that separate test batteries might be misadministered. To prevent such an occurrence, a single battery containing the union of tests developed for all three jobs was assembled and administered to corrections officers as well as to probation officers and juvenile counselors.

Each test in Booklet 3 contained 35 items (175 items in all). A time limit of 60 minutes was set for the entire booklet. This allowed ten minutes per test (twice the estimated time necessary to complete 35 items) and left an additional ten minutes of testing time. Incumbents were free to divide the 60 minutes among the five tests as they saw fit. However, there were prominent warnings that no more than ten minutes should be spent working items for any given test until all tests had been attempted.

Noncognitive Tests

Noncognitive marker scales likely to predict job performance were identified at the same time and using the same methods as described for the cognitive marker tests. Four of the five tests thus identified are shown in Table 3.8. The publisher of each test was contacted to secure permission to use their test in this project and to ascertain whether the Board of Corrections would be granted permission to score the tests. For one of the marker tests (Guilford-Zimmerman Temperament Survey) it seemed unlikely that the necessary permissions would be obtained. The Guilford-Zimmerman Temperament Survey was omitted from the experimental selection battery.

Although not identified as a potentially useful scale in the analysis conducted to identify noncognitive marker tests, the Correctional Officers' Interest Blank (COIB) was selected for inclusion in the experimental selection battery based on its proven performance in similar studies. The additional noncognitive tests (Multidimensional Personality Questionnaire Control and Hogan's Empathy) targeted for the other positions were also included in the battery.

With the exception of 18 items from the COIB, all of the items from the noncognitive tests were in true/false format. These items were edited to change all occurrences of the pronoun "I" to "you". These changes necessitated other minor editorial changes so that the items read well. A few items were omitted because of their invasive nature. These items were potentially self-incriminating (e.g., "You have never been in trouble with the law") or dealt with topics too personal for use in an employment setting (e.g., "You have been in trouble one or more times because of your sex behavior"). Editing was kept to a minimum. STC staff members also reviewed the items for biased or objectionable items, leading to the deletion of five items, as well as several minor editorial changes in item content.

All noncognitive items were assembled into one booklet with two sections. The first section contained the 228 true/false items from the noncognitive tests. Items from the different scales were mixed together unsystematically. The second section contained the remaining COIB items. These items required incumbents to pick from among three alternatives the statement they would like most and the one they would like least. Based on an estimated response rate of five items per minute (12 seconds per item) a time limit of one hour was set for the noncognitive items booklet.

Sample Items Booklet

To provide the incumbents in the validation sample an opportunity to become familiar with the cognitive ability items in the experimental selection battery, a booklet with illustrative items from each cognitive ability test was assembled. For each test the purpose of the examination was briefly explained, two or three sample items given with the correct answers explained, and eight to ten practice items (with answers) given. This booklet was distributed to all incumbents tested and was the prototype of a similar booklet to be made available to future job applicants as a preview of the Corrections Officer Examination.

CHAPTER FOUR: CRITERION DEVELOPMENT

A criterion-related validation strategy was used in this study. In a criterion-related approach, tests are valid if test scores can be demonstrated to be statistically related to a reliable measure of job performance. The agencies participating in this study used different methods of assessing the performance of their incumbents. Thus, it was not possible to use existing performance ratings as the criterion for judging the validity of the experimental tests. Even if a standard method of assessing incumbent performance had been available, it would not have proved useful because professional standards mandate that criterion-related studies include a relevant and reliable criterion measure. Existing job performance data were of unknown relevance and unknown reliability. Thus, it was necessary to develop new measures of job performance to be used as the criterion for assessing the validity of the predictors.

To ensure the relevance of the criterion measures, job analysis data were used to guide criterion development. These data provided a clear picture of the most important aspects of the corrections officer job and thus provided an ideal starting point for criterion development. The goal of criterion development was to create a series of scales that fully cover the relevant aspects of the corrections officer job. These scales would then be used by supervisors to rate the job performance of the incumbents who completed the experimental selection battery.

A critical incidents approach was taken to criterion development. In this approach, corrections officer supervisors wrote behavioral incidents representing examples of poor, average, and excellent performance for critical aspects of the corrections officer job. These incidents, when analyzed and summarized, provided the benchmarks that were used by other supervisors to rate the performance of incumbents in these important areas. Details of the critical incident methodology are given below.

Collecting Critical Incidents

Supervisors were recruited to participate in this phase of the study in such a way that large and small counties from both northern and southern California were represented. Participating supervisors were intimately familiar with the duties of corrections officer incumbents and were articulate and able to express themselves well in writing. A total of 34 supervisors attended 1/2-day workshops conducted by a PDI psychologist. During the workshops, the critical incident process was explained and the critical incidents written.

To guide their thinking about the job, supervisors were given a summary of the job analysis consisting of descriptions of the task categories developed for use in linking worker characteristics to task categories (See Chapter Three).

The task categories were explained to the supervisors who were then provided forms (See Figure 4.1) on which to record their critical incidents. These forms gave general instructions about writing critical incidents and provided space for the supervisors to: 1) describe the circumstances leading up to each incident; 2) describe what the incumbent did during the incident; 3) rate the incumbent's performance on a 1 (extremely ineffective) to 9 (extremely effective) point scale; and 4) indicate into which (if any) of 10 general task categories the incident fit.

Each supervisor was able to generate approximately 10 incidents yielding a total of 343 critical incidents from the 34 supervisors. These incidents provided the raw material from which development of the criterion measures proceeded.

Revising and Retranslating the Critical Incidents

The incidents were edited to emphasize the behaviors involved, to separate double incidents, and to ensure that all incidents were described in a similar format. PDI staff members rationally grouped the edited incidents into categories based upon the behaviors involved in each incident. The goal was to create categories which included only those incidents revealing varying levels of effective performance for similar or related job tasks. Ten such categories of incidents were created and are described in Table 4.1. To corroborate the categories and to ensure that each incident was correctly assessed as to its degree of effectiveness, the 343 incidents were "retranslated" by corrections officer supervisors. During retranslation a second set of supervisors was given a list of the edited incidents and descriptions of the 10 categories. The supervisors were asked to place each incident in the most appropriate category. Retranslation was necessary to be sure that from the perspective of persons very familiar with the job, each incident was correctly grouped and to ascertain the degree of effectiveness each incident represented. For reasons of administrative feasibility, the 343 incidents were assembled into two booklets, each with 176 incidents (the booklets overlapped by nine incidents to permit the data collected for each booklet to be equated). The booklets contained detailed instructions for the retranslation task and provided descriptions of the 10 performance categories to be used in making the ratings. Each booklet contained incidents believed to belong in each of the 10 categories.

Figure 4.1

PERFORMANCE INCIDENT FORM

Thinking back, focus your attention on any specific thing that you saw a person do that made you think of him/her as a highly effective worker, an average worker, a very ineffective worker, or any shade of effectiveness in between. In other words, think of a specific example of job behavior that gave you some feeling for how well the individual was doing his/her job. (Please do not record the names of the people involved.)

1. What were the circumstances leading up to the incident?

2. What did the individual do that made you think he/she was a good, average, or poor performer?

3. How effective do you feel this example of performance was? (Circle a number.)

1	2	3	4	5	6	7	8	9
extremely ineffective		ineffective		about average		effective		extremely effective

4. What, if any, general category would the above action fit into?

Table 4.1

Corrections Officer Performance Dimensions Used During Retranslation

---

A. MAINTAINING SECURITY

Maintaining the security of the facility and inmates: conducting counts properly and efficiently; keeping doors, etc. locked; operating doors from control; maintaining the security of facility keys; ensuring that inmates are where they should be and not where they shouldn't be; verifying the identity of inmates, for example, before releasing, and verifying the identity of visitors.

B. SEARCHING

Using effective search procedures; inspecting food, mail, packages for contraband; searching cells, dorms, halls, yard, and other facility areas for contraband and taking appropriate action when found.

C. ALERTNESS TO POTENTIAL PROBLEMS

Being alert to signs of potential danger or disturbance; noticing sounds or subtle changes in inmate behavior or conversation that might indicate some kind of trouble, and taking the appropriate steps to prevent the trouble; investigating suspicious activity or situations; recognizing signs of suicidal tendency, insulin shock, etc. and taking appropriate action.

D. KEEPING/CHECKING RECORDS

Performing all paperwork necessary for the correct and efficient operation of a facility: reviewing court documents, figuring release dates correctly; running warrant checks; logging daily activities or incidents; logging inmates' property and money, and correctly returning same on release; writing reports, such as escape reports, and others, that are clear, complete, and concise.

E. USE OF FORCE/RESTRAINTS

Using the appropriate physical force, restraints (cuffs, etc.) or weapon for the situation; skillfully applying self-defense techniques when necessary; using the proper procedure to subdue inmates; recognizing when force is required and using proper amount (being careful not to use excessive force when it is inappropriate).

Table 4.1 (continued)

F. HANDLING EMERGENCIES/CRISES

Remaining calm and taking appropriate actions in emergency or crisis situations such as fire, injury, suicide attempt, escape, or fighting inmates; sounding or responding to an alarm with the proper procedure.

G. DEALING WITH HARASSMENT

Responding calmly and appropriately to verbal harassment or uncooperativeness; maintaining self-control in the face of harassment or intimidation; not responding to harassment with threats of violence or foul language, but with reasoning and patience, and taking any appropriate disciplinary actions.

H. INTERACTING WITH INMATES

Displaying a concern for inmates' health and general well-being; showing respect for inmates' rights and privileges; treating inmates fairly and consistently; communicating effectively with inmates, gaining their cooperation; listening and responding, when appropriate, to inmates' questions, concerns, complaints; following through on inmate requests and helping them work out problems when appropriate.

I. INTERACTING WITH CO-WORKERS

Getting along with co-workers, supervisors, and other criminal justice personnel; communicating effectively with other officers; keeping others completely informed of things they should know; assisting/supporting other officers when needed or when requested to do so; helping train new officers.

J. COMMITMENT TO JOB

Demonstrating commitment to the job and respect for a facility's policies and procedures: putting in more time and effort than required versus just doing enough to get by on the job; doing things to improve the operation or the "image" of a facility on one's own time and/or initiative versus not pulling one's fair share of the work load, for example, by sleeping on the job or abusing sick leave.

Thirty-five corrections officer supervisors were selected to categorize the incidents into one of the 10 possible job performance categories and to rate the degree of effectiveness each incident represents. Effectiveness ratings were made on a 9-point scale. Nineteen supervisors completed the first booklet and 16 completed the second. These supervisors represented 9 agencies and were responsible, on average, for supervising 17.2 corrections officer incumbents. They had been supervisors for an average of 5.4 years and the 34 who had also served at corrections officer incumbents reported an average of 5.7 years experience.

Developing Behavioral Summary Scales

For every retranslated incident, the proportion of supervisors who placed the incident in each category was computed, as was the mean effectiveness rating and the standard deviation of the ratings about that mean. These data were used to revise the categorization and to select incidents to appear as behavioral examples of poor, adequate, and excellent performance. Based upon the retranslation data, the category "Dealing with Harassment" was dropped and its items reassigned to the "Handling Emergencies/Crises" and "Interacting with Inmates" categories. This was necessary because only 11 incidents were categorized as "Dealing with Harassment" -- too few incidents to permit construction of a separate rating scale.

Incidents within each category were sorted into those indicating poor, adequate, and excellent performance and used to construct rating scales. Each scale (See Figure 4.2) provided a description of the performance category to be rated along with a 7-point rating scale. Ratings of 1 or 2 were labeled as "needs improvement", ratings of 3, 4, or 5 were labeled "fully adequate", and ratings of 6 or 7 were labeled "excellent". Three to five incidents were abstracted to provide examples of the kind of behaviors that should result in a given performance rating. Instructions to raters were developed to emphasize that the behavioral examples were not checklists, but rather provided to guide the raters' thinking about typical kinds of behaviors and therefore the level of performance of each incumbent being rated.

Other Ratings Scales

Worker Characteristic Ratings. In addition to the 9 behavioral summary scales, 22 of the 24 worker characteristics created during predictor development (see chapter three) were transformed into 7-point rating scales. Ratings of 1 or 2 indicated "low" standing (among the lowest one-third of employees), 3, 4, or 5 indicated "average" (in the middle one-third), and ratings of 6 or 7 indicated a "high" standing (upper one-third) on that worker characteristic.

Figure 4.2
Sample Behavioral Summary Scale

A. MAINTAINING SECURITY

Conducting counts properly and efficiently; keeping doors, etc. locked; maintaining the security of facility keys; ensuring that inmates are supervised and that they are in the proper place; verifying the identity of inmates and visitors.

1 2	3 4 5	6 7
Needs Improvement	Fully Adequate	Excellent
. Does not follow proper procedures for taking counts; takes counts inaccurately or fails to take counts.	. Follows proper procedures for taking counts; takes counts accurately and takes appropriate action if inmates are missing.	. Promptly discovers if an inmate is missing, by taking counts frequently and accurately, making sure to "see skin"; may randomly initiate a second count to ensure accuracy.
. Leaves security doors unlocked or open; doesn't secure the area before opening security doors from control.	. Routinely checks to make sure doors are locked; ensures that keys are secure; notices and takes proper action when keys are missing.	. Makes careful checks of security doors, notices locks which can be "pooped" or easily forced open, and takes proper action.
. Leaves keys lying out in the open; misplaces or loses keys.	. Does not leave inmates unsupervised; makes sure that inmates returning from outside the facility go to the correct housing unit; checks the identity of inmates prior to releasing them, and checks the identity of visitors.	. Prevents inmates from escaping or being accidentally released due to mistaken identity, by very carefully checking their photographs and/or fingerprints.
. Leaves inmates unsupervised; doesn't verify the identity of inmates before allowing them to enter housing unit or before releasing them; doesn't make sure that visitors are who they say they are.		

The worker characteristics of "Psychology Training", "Driving", and "Using Firearms", were not included as job performance ratings scales because supervisors were unlikely to be able to rate incumbents' standing in those worker characteristics reliably or because ratings were not appropriate methods of assessing incumbents' standing for those characteristics. The worker characteristic "Quantitative Skill" was renamed, "Arithmetic Skill" so that it more accurately reflected the type of abilities being assessed. Similarly, "Tolerance for Supervision and Structure" was renamed "Non-autonomy" and the worker characteristic dealing with vision and hearing was separated into two scales: one for vision and one for hearing.

Overall effectiveness. Finally, an overall effectiveness scale was constructed on which supervisors were asked to rate how well an incumbent met the standards and expectations of the job compared to other individuals with similar levels of experience. Ratings were made on the same 7-point scale ("Needs Improvement" to "Excellent") used for the behavior summary scales.

The final criterion rating scales consisted of 9 scales developed from the critical incidents data, 22 scales based upon the worker characteristics, and a single, overall effectiveness scale.

CHAPTER FIVE: DATA COLLECTION

Once the experimental selection battery and criterion ratings materials had been developed and assembled, the tests were to be administered to a sample of corrections officer incumbents and job performance ratings collected from the incumbents' superiors. Following are the efforts undertaken to secure a sample of corrections officer incumbents of sufficient size and composition to permit an accurate and efficient assessment of the validity of the experimental selection battery.

Predictor Data Collection

Predictor data collection began with the setting of targeted sample sizes for each of the participating agencies. Sample sizes were based upon relative agency size; however, smaller agencies were requested to test a proportionally larger number of their incumbents than were larger agencies. The initial sampling plan called for testing 497 corrections officers from 16 sheriff's departments.

Department coordinators were contacted by phone to secure permission to test the requested number of corrections officers. As agency participation was voluntary, a certain amount of negotiation was involved in arriving at an agreed-upon number of incumbents to be tested in each agency. In spite of the heavy and repeated requests for department resources imposed by this project, department coordinators were universally willing to go the "extra mile" to ensure that testing ran smoothly.

Information about the testing process was sent to each coordinator outlining the purpose of testing and reviewing the procedures to be involved. An important part of the information sent to the coordinators concerned the method by which incumbents were to be selected for testing. Because job performance ratings were to be collected for all incumbents tested, a selection method was devised that attempted to minimize the burden put upon each agency's supervisory staff. This sampling method called for the ordering of supervisors based upon their experience. Incumbents were then chosen by selecting all of the incumbents of the most experienced supervisor, all of the incumbents of the second most experienced supervisor, and so on until the targeted number of incumbents had been selected.

Testing was scheduled to occur during June 1984 but various organizational problems led to a number of delays so that testing was not completed until November of that same year. The first

problem to occur was the withdrawal of a participating agency, thereby reducing the targeted sample from 497 to 251. In order to offset this loss, PDI contacted other participating agencies in an attempt to secure additional participants. This effort increased the sample size by 39 persons. An additional step of contacting agencies that had participated during the job analysis phase but that were not involved in testing was undertaken. Of these requests for participants, three agencies were willing to commit a total of 45 additional participants to the validation phase of the project.

At that point, the maximum possible sample size for the validation phase stood at 335. However, it was not reasonable to assume that 335 usable test batteries would be completed. A more reasonable estimate, allowing for persons who failed to appear for testing, etc., would be that approximately 85% of the batteries would be completed. Thus, the expected sample size stood near 285. In order to secure additional incumbents, STC consultants contacted three agencies that had not previously participated in any aspect of the project. These three agencies supplied an additional 55 incumbents, raising the total targeted sample to 350 and the expected sample to approximately 332.

By October of 1984, phone canvassing of the participating agencies made it clear that no additional incumbents were available for testing and that all of the incumbents who were going to be tested had been tested. A count of the data collected to that point revealed a sample size of 289. When the number of incumbents in protected groups was counted, it appeared that too few black or Hispanic corrections officers had been tested to permit a reasonable evaluation of the fairness of the selection procedures. STC consultants then arranged with another sheriff's department for the testing of 50 additional black or Hispanic corrections officers bringing the total number of incumbents for whom predictor data were collected to 339. Testing of the 50 additional corrections officers was contingent upon PDI paying $75 per participant ($3,750) to offset the overtime payments incurred as a result of testing those incumbents. In addition, a PDI staff member flew to California to oversee testing of these additional incumbents. The testing of the 50 additional incumbents occurred just prior to Thanksgiving and marked the end of the predictor data collection efforts of this study.

Criterion Data Collection

Soon after the experimental selection batteries were shipped to coordinators, job performance rating materials were also shipped in sufficient quantity to permit supervisors to rate the performance of all incumbents scheduled to be tested. These materials consisted of a performance appraisal booklet and two response booklets. The performance appraisal booklet introduced the performance rating task

and provided general instructions about the rating process as well as detailed instructions for the completion of each rating scale. Each response booklet contained space to rate the performance of five incumbents. Each supervisor also received a cover memo providing more information about the ratings task (e.g., where it fit into the overall research effort), enlisting the supervisor's assistance in maintaining the confidentiality of the ratings he/she made, and directing the supervisor to the department coordinator in the event that questions about the ratings process arose.

Job performance ratings were returned by a total of 106 corrections officer supervisors from 18 agencies. These raters reported having been employed as supervisors for an average of 4.2 years. Four of the supervisors were female, the rest (102) were male. Nine supervisors were black, two were Native Americans, 7 were Hispanic, and 84 were white. About half of the supervisors reported that they worked in maximum security facilities with the remaining half fairly evenly distributed among minimum security, medium security, and "other" facilities. These 106 supervisors rated the performance of 282 incumbents on each of 9 behavioral summary scales, 22 worker characteristic scales, and the overall effectiveness scale representing a total of 9,024 ratings.

As predictor and criterion data arrived, checks were made to ensure that every correction officer incumbent tested had also been rated. In those cases where performance ratings did not arrive for a tested incumbent, department coordinators were contacted and asked to follow-up with the appropriate supervisor to ensure that the ratings were made. Despite frequent requests to follow-up on ratings collection, it was not possible to obtain job performance ratings for every corrections officer who completed the experimental selection battery.

Final Validation Sample

The criterion-related validation strategy employed in this study relied upon job performance ratings as the yardstick against which the validity of the predictors was to be judged. The final validation sample (i.e., those corrections officer incumbents whose data were analyzed to determine the validity of the predictors) consisted of those persons for whom <u>both</u> usable predictor and usable criterion data were collected. As was note, 339 corrections officers were tested, but only 282 had their job performance rated by their supervisor. Thus the maximum possible number of incumbents for whom both predictor and criterion data were available was 282. Unfortunately, there were other ways incumbent data were lost from the final validation sample. The missing data problems encountered in this project and the impact of those problems upon the size of the final validation sample are discussed in the following. Table 5.1 lists, by county, the number of incumbents tested and the number lost to the validation sample for the reasons enumerated. Following this discussion, the sample demographics are presented and the technical adequacy of the sample defended.

Table 5.1

Composition of the Final Validation Sample

County	Number of COs Tested	Lost from Validation Sample: Predictors w/o Name	COs not Rated	Low Tenure	Not Well Known	Final N
Fresno	26	-	-	-	3	23
Imperial	15	-	5	-	-	10
Los Angeles	51	-	1	2	2	46
Merced	7	2	5	-	-	0
Monterey	18	-	-	5	-	13
Napa	5	-	5	-	-	0
Nevada[a]	3	-	-	-	-	3
Riverside[a]	15	-	-	1	-	14
Sacramento[a]	10	-	-	5	-	5
San Diego[a]	29	-	-	1	-	28
San Francisco	38	34	1	-	-	3
San Mateo	4	-	1	-	-	3
Santa Barbara	14	-	1	-	-	13
Santa Cruz	11	-	-	-	-	11
Shasta	13	-	-	-	-	13
Siskiyou	0	-	-	-	-	0
Sonoma	11	-	-	2	-	9
Stanislaus	9	-	-	-	-	9
Ventura	50	-	-	6	5	41
Yolo	11	-	2	-	-	9
			Final Validation Sample			253

[a]Not among the initially targeted agencies.

Missing Name. Incumbents' names were used to match predictor and ratings data. Incumbents who failed to put their names on their answer sheets prevented this match from being made for their data. Missing names were a major problem for only one county where 89% of the incumbents tested did not put their names on the test answer sheets. A total of 36 cases were lost due to missing names.

Missing Ratings. A second way in which cases were lost from the validation sample was through missing ratings data. Despite repeated efforts to follow-up with department coordinators, ratings were not obtained for 21 incumbents who had completed the experimental selection battery.

Low Tenure. The validation sample was intended to consist of only experienced incumbents (i.e., those on the job for at least six months). The job performance of incumbents with six or fewer months experience would be difficult to rate because supervisors may not have had an adequate opportunity to observe these incumbents' performance under a variety of job conditions. In order to ensure that the job performance ratings of all incumbents accurately assessed their ability to perform the job, incumbents with six or fewer months tenure were not included in the final validation sample.

Inaccurate Ratings. When rating the performance of their incumbents, supervisors were also asked to indicate how well they knew the incumbent whose performance they were assessing. Incumbents whose ratings were made by supervisors who said that they did not know that incumbent "well at all" (i.e.. rating of 4) were dropped from all analyses. In order to judge the validity of the predictors, an incumbent's true level of performance must be known. Obviously, the ratings of an incumbent not at all well known by his/her supervisor could not be an accurate measure of the incumbent's true job performance.

Demographics of Final Validation Sample

As a result of the data losses noted, the final validation sample consisted of the 253 corrections officer incumbents for whom accurate performance ratings were available. The demographic composition of the final sample is detailed in Tables 5.2 - 5.4. The sample is primarily male and approximately one-half of the corrections officers are members of protected groups (principally, blacks or Hispanics). Most of the corrections officers in the final sample have some form of post high school education (i.e. Associate Degree, B.S., or B.A. Degree) and more than one-half of the sample have been corrections officers for two or more years (only 15% have less than one year's experience on the job).

Table 5.2

Gender and Ethnic Status of Final Validation Sample[a]

Group	Number	Percentage
Gender		
Male	194	77
Female	58	23
Ethnic Status		
Black	44	18
Native American	5	2
Asian	4	2
Hispanic	56	23
White	138	55
Other	2	1

[a]Totals may vary due to missing data.

Table 5.3

Education and Tenure of Final Validation Sample[a]

Group	Number	Percentage
Education		
High School	67	29
Tech./Vocational	40	17
Associate Degree	79	34
B.A./B.S. Degree	41	17
Some post-graduate	6	3
Post-graduate Degree	2	1
Tenure		
Less than 3 months	b	
3-6 months	b	
6 months to 1 year	40	17
1 to 2 years	60	25
2 to 5 years	104	44
More than 5 years	35	15

[a]Totals may vary due to missing data.

[b]COs with fewer than 6 months tenure were not included in the final validation sample.

Table 5.4

Relationship of Incumbents to Supervisors in Final Validation Sample[a]

Group	Number	Percentage
Length of Time Supervised		
Less than 3 months	8	3
3 to 6 months	56	22
6 months to 1 year	83	33
1 to 2 years	59	24
2 to 5 years	40	16
More than 5 years	4	2
How Well Known by Supervisor		
Very well	118	47
Pretty well	116	46
Not too well	18	7
Not at all well	b	

[a]Totals may vary due to missing data.

[b]Incumbents known "not at all well" were dropped from the final validation sample.

Technical Adequacy of Sample

One concern in conducting a criterion-related validation study is whether enough incumbents can be tested to permit a technically adequate evaluation of the validity and fairness of the predictors. Several analyses were performed to determine the technical adequacy of the sample collected in this study (See "Report on the Technical Adequacy of the Final Validation Sample for Corrections Officers" available from the Board of Corrections). The major issue was whether the sample size provided sufficient statistical power to detect predictor validity and subgroup differences should they have existed. The analyses uniformly demonstrated the adequacy of the validation sample for all of the analyses performed during the course of the study. Thus, the 253 corrections officer incumbents for whom predictor and job performance data were available provided a more than sufficient basis for judging the validity of those predictors as measures of corrections officers' job performance.

The analyses reported in the following chapters were based upon all available data. This ensured that the results of those analyses were as stable as possible. As a consequence, the sample size varied slightly from analysis to analysis. This occurred because an incumbent who was missing data essential to one analysis might have had all the data required for another. Thus, the sample size would have varied because the incumbent was excluded from the first analysis but included in the second. Exact sample sizes for the analyses are reported in the tables summarizing the analyses. Slight variations in sample size are to be expected for the reason given above.

CHAPTER SIX: ANALYSES AND RESULTS

Many different analyses of the predictors and criteria were conducted to document their validity and fairness. The goal of all analyses was to obtain a combination of predictors that simultaneously:

1) would validly predict job performance;

2) was fair (i.e., did not unfairly impact protected groups); and

3) was administratively practical for all agencies participating in the STC Program to implement.

The analyses conducted to meet the above goals can be divided into three categories: criterion analyses, predictor analyses, and validation and fairness analyses.

Criterion analyses included those necessary to determine: a) that the job performance ratings were reliable; b) how best to combine the 32 rating scales into a single measure of job performance; and c) whether there were specific aspects of performance differentially predictable by one or more of the 15 tests.

Predictor analyses included those necessary for final revision of the tests (especially Reading and Observational Memory) to: a) document the psychometric quality of the tests; and b) describe the level of performance obtained by the corrections officers in the final validation sample.

Validation and fairness analyses were those necessary to determine: a) how best to combine the tests from the experimental selection battery into a useful, valid predictor of job performance; and b) that the selected tests did not unfairly discriminate against candidates from protected groups.

While each of the above categories of analysis are conceptually distinct, there was considerable overlapping of analyses. Each category will be described separately although occasional reference to analyses from other categories will be necessary.

Criterion Analysis

The criterion analyses had as their basic goal the development of a reliable measure of overall job performance. This criterion would serve an important role in later validation analyses. An additional and equally important goal of the criterion analyses, was to discover whether there were specific aspects of job performance that could be better predicted by single tests than by the Corrections Officer Examination. The presence of such differentially predictable aspects of job performance would have broadened the usefulness of the validation research as it would have resulted in a "menu" of selection instruments from which agencies could pick according to their needs.

Data collection and cleaning. Incumbents' job performance was rated by their supervisors on 9 behavioral summary scales, 22 worker characteristic scales, and on a single, overall effectiveness scale (see Chapter Four). To preserve the confidentiality of the performance ratings, supervisors were provided pre-addressed envelopes for mailing their rating materials directly to PDI. No one other than PDI staff members and the supervisor making the ratings had access to the performance ratings of any individual corrections officer.

When the ratings arrived at PDI, they were reviewed for completeness and assigned an identification number. Ratings were then keypunched. The keypunchers identified any ratings where there was a question as to the intended rating. PDI staff reviewed all ratings so identified and determined the correct ratings value or set that value to "blank", indicating a missing value.

The mean rating value for each of the 32 individual rating scales is shown in Table 6.1. As can be seen, corrections officers varied in the ratings they received on the 32 scales. For example, performance in the area of Searching is more highly rated than is performance in the area of Street Knowledge. It should be noted that comparisons of this sort are only appropriately made between ratings from similar scale types (i.e. worker characteristic to worker characteristic or behavioral summary scale to behavioral summary scale) and are valid only for average group performance. To compare an individual's ratings across scales, it would first be necessary to convert the ratings into standardized units. This is necessary because a rating of 4 on a scale with an average of 4.5 represents poorer relative performance than does a rating of 3 on a scale with an average of 2.5.

The overall performance of corrections officers in the final validation sample ranges from "needs improvement" to "excellent". The mean rating for the overall effectiveness scale was 4.75 with a standard deviation of 1.12 indicating that 90% of all corrections officers received

Table 6.1

Descriptive Statistics for 32 Job Performance Ratings Scales

Scale	Number of COs Rated	Avg.[a]	Stnd. Dev.
Behavioral Summary Scales			
Maintaining Security	249	4.75	1.15
Searching	242	4.68	1.17
Alertness to Potential Problems	249	4.78	1.21
Record Keeping/Report Writing	247	4.37	1.24
Use of Force/Restraints	241	4.52	1.13
Handling Emergencies/Crises	247	4.81	1.20
Interacting with Inmates	249	4.55	1.29
Interacting with Co-workers	250	4.83	1.24
Commitment to Job	250	4.84	1.26
Worker Characteristic Scales			
Facility-Specific Knowledge	249	4.91	1.26
Knowledge of the Legal System	248	4.20	1.22
Street Knowledge	245	3.96	1.35
First Aid	247	4.63	1.05
Operating Communications Equipment	248	4.75	1.01
Oral Communications	249	4.68	1.17
Reading	249	4.72	1.02
Written Communications	248	4.41	1.18
Arithmetic Skill	230	4.80	1.06
Observational Activity	250	4.67	1.11
Self-Defense/Physical Restraint of Others	242	4.65	1.23
Social Skills	249	4.64	1.15
Motivation/Dependability	250	4.78	1.36
Stress Tolerance	250	4.69	1.22
Dominance/Self-Esteem	250	4.53	1.26
Tolerance for Supervisor and Structure	250	4.86	1.24
Flexibility	250	4.58	1.23
Tolerance of Negative Job Aspects	250	4.52	1.18
Searching	243	4.91	1.15
Ability to Meet Visual Requirements	250	5.00	1.13
Ability to Meet Hearing Requirements	250	4.94	1.12
Ability to Meet Physical Requirements	250	4.82	1.33
Overall Job Performance	228	4.75	1.12

[a]Ratings made on a 7-point scale.

ratings between 3 and 7 on the 7-point scale. The distribution of scores is skewed somewhat toward higher ratings and, as a result, is slightly leptokurtic (peaked).

Inspection of the intercorrelations of the 32 ratings scales indicated that ratings within categories were not more highly interrelated than were ratings across categories. For example, the average intercorrelation of the 9 behavioral summary scales is .47 and the average intercorrelation of the 22 worker characteristic scales is .45. The average correlation of the 9 behavioral summary scales with the 22 worker characteristic scales is .42 indicating that neither of these rating formats contributed significant unique method variance to the ratings data. This finding allows the ratings to be combined across scales as is discussed in the next section.

Developing an overall criterion. The principal goal of the criterion analyses was to develop a single measure of job performance that would serve as the criterion for subsequent validation analyses. A criterion-related research design requires that there be a single criterion against which the validity of the overall selection battery is judged.

Inspection of the data in Table 6.1 indicated no reason to give precedence to any one of the individual rating scales as a measure of overall effectiveness. By itself, the 7-point overall effectiveness scale was not appropriate as the principal criterion for two reasons. First, individual scales do not provide as reliable measures as do combinations of scales, nor do they provide sufficient variation in criterion scores to permit an accurate assessment of the magnitude of the relationship between predictor and criterion. Second, the overall effectiveness scale may have meant different things to different raters. For example, rater A may have placed more weight on Oral Communications while rater B may have weighted Vision and Hearing more heavily. Because the overall effectiveness scale was inappropriate and because no other scale stood out as deserving special attention, an overall effectiveness score was constructed by averaging across all 32 rating scales.

The overall effectiveness criterion (hereafter called overall criterion) was computed by: 1) standardizing each rating scale (T-scores; mean = 50.00, SD = 10.0); 2) computing an average rating for each scale type (i.e., behavioral summary scales, worker characteristic scales, and overall rating) and; 3) taking the average rating across the three averages. As long as an incumbent received ratings on 90% or more of the 32 scales, missing ratings were treated by averaging across those scales for which ratings were made. Standardizing the scales prior to computing the overall criterion, resulted in each type of scale contributing equally to the final average.

This method resulted in an overall criterion with a mean of 50.3 and a standard deviation of 7.5 (based upon 253 cases). Table 6.2 presents the correlation of this overall criterion with each of the individual rating scales. The average of the 9 behavioral summary scales correlated .95 with the overall criterion, and the average of the 22 worker characteristic scales correlated .92 with the same criterion. The magnitude of these correlations is due in part to the fact that the behavioral summary and worker characteristic scales each comprise one-third of the overall criterion. Thus, the correlations are inflated by the fact that they, in large part, represent the relationship of an average with itself. The average of the behavioral summary scales correlated .84 with the average of the worker characteristic scales. The average of the behavioral summary scales and the average of the worker characteristic scales correlated .83 and .77, respectively, with the overall effectiveness scale.

An internal consistency approach was taken to compute the reliability of the overall criterion. Criterion reliability was found to be .96. Originally it was intended that multiple ratings would have been collected for each incumbent in the validation sample. However, discussions with department coordinators indicated that it would be very difficult to locate even two supervisors who were sufficiently familiar with an incumbent's performance to provide accurate ratings.

The internal consistency approach used is the same that was described in Chapter 3 to estimate the reliability of other ratings data. This approach is somewhat susceptible to overestimating the ratings reliability due to a tendency on the part of raters to form a global impression of incumbents' performance and not to distinguish between different levels of performance on the different ratings scales (halo error). The rating instructions warned raters against this tendency and requested that care be taken in making the ratings. Nevertheless, the reliability reported above will slightly overestimate the true reliability of the ratings.

Ratings reliability is important in validation research only because it places a limit on the validity of the predictors. The more unreliable the ratings, the lower the maximum possible validity. Reliability estimates are commonly used to correct validity coefficients for criterion unreliability. Such corrections always increase the magnitude of the estimated validity, with the degree of increase being inversely proportional to the magnitude of the reliability estimate. Because the method used to estimate the reliability of the ratings slightly overestimates that reliability, it is a conservative estimate. That is, its use to correct the validity estimate for criterion unreliability will not cause a large increase in the magnitude of the validity coefficient.

Table 6.2

Correlation of 32 Job Performance
Ratings Scales with the Overall Criterion

Scale	Corr. w/overall criterion[a]
Behavioral Summary Scales	
Maintaining Security	.74
Searching	.62
Alertness to Potential Problems	.68
Record Keeping/Report Writing	.66
Use of Force/Restraints	.64
Handling Emergencies/Crises	.77
Interacting with Inmates	.70
Interacting with Co-workers	.71
Commitment to Job	.70
Worker Characteristic Scales	
Facility-Specific-Knowledge	.67
Knowledge and Legal System	.61
Street Knowledge	.56
First Aid	.48
Operating Communications Equipment	.59
Oral Communications	.70
Reading	.59
Written Communications	.63
Arithmetic Skill	.53
Observational Activity	.76
Self-Defense/Physical Restraint of Others	.63
Social Skills	.73
Motivation/Dependability	.69
Stress Tolerance	.66
Dominance/Self-Esteem	.71
Tolerance for Supervisor and Structure	.72
Flexibility	.70
Tolerance of Negative Job Aspects	.68
Searching	.69
Ability to Meet Visual Requirements	.51
Ability to Meet Hearing Requirements	.48
Ability to Meet Physical Requirements	.51
Overall Job Performance	.94

[a]The number of cases varies from approx. 220-250; all r's are significant at alpha = .05

Search for additional criteria. Apart from the development of an overall criterion, the criterion analyses had as an additional goal the development of criteria relating to more specific aspects of job performance. To the extent that such specific criteria were predictable using one or more of the tests in the experimental selection battery, agencies needing to assess those specific areas of job performance would have valid instruments at their disposal.

The search for additional criteria was conducted via factor analysis of the ratings data. The ratings intercorrelation matrix was divided into three submatrices: one containing the intercorrelations of the behavioral summary scales, the second containing the intercorrelations of the worker characteristic scales, and a third containing the intercorrelations of the behavioral summary and the worker characteristic scales. None of the submatrices contained the overall effectiveness rating scale. An iterative, principal components analysis, followed by varimax rotation to simple structure, was performed for each submatrix as is further described.

Analysis of each submatrix began with the extraction of a single factor. This was followed by additional analyses extracting two, three, . . . N, factors at each step. After each additional factor was extracted, the factor loading matrix was examined to determine whether the resulting solution made sense in light of what was known about the corrections officer job. Additionally, the traditional factor solution criteria of eigenvalue magnitude and variance accounted for were monitored. The iterative procedure continued until the extraction of the next factor either made no sense, or drastically exceeded typical rules of thumb for acceptable factor solutions (e.g., eigenvalue less than 1.0).

A two-factor solution was determined to be most appropriate for the behavioral summary scales submatrix and a five-factor solution was accepted for the worker characteristics submatrix. As these two solutions were judged to have considerable overlap, the factor analysis involving both behavioral summary and worker characteristic scales was taken as the most appropriate basis on which to construct additional, specific criteria. Descriptions of the resulting factors are given in Table 6.3. Figure 6.1 indicates how each of the 31 behavioral summary or worker characteristic scales loaded on the five-factor solution. It should be noted that the "all or nothing" loadings shown in Figure 6.1 represent the forced simple structure solution that was used to compute scores for each additional criterion.

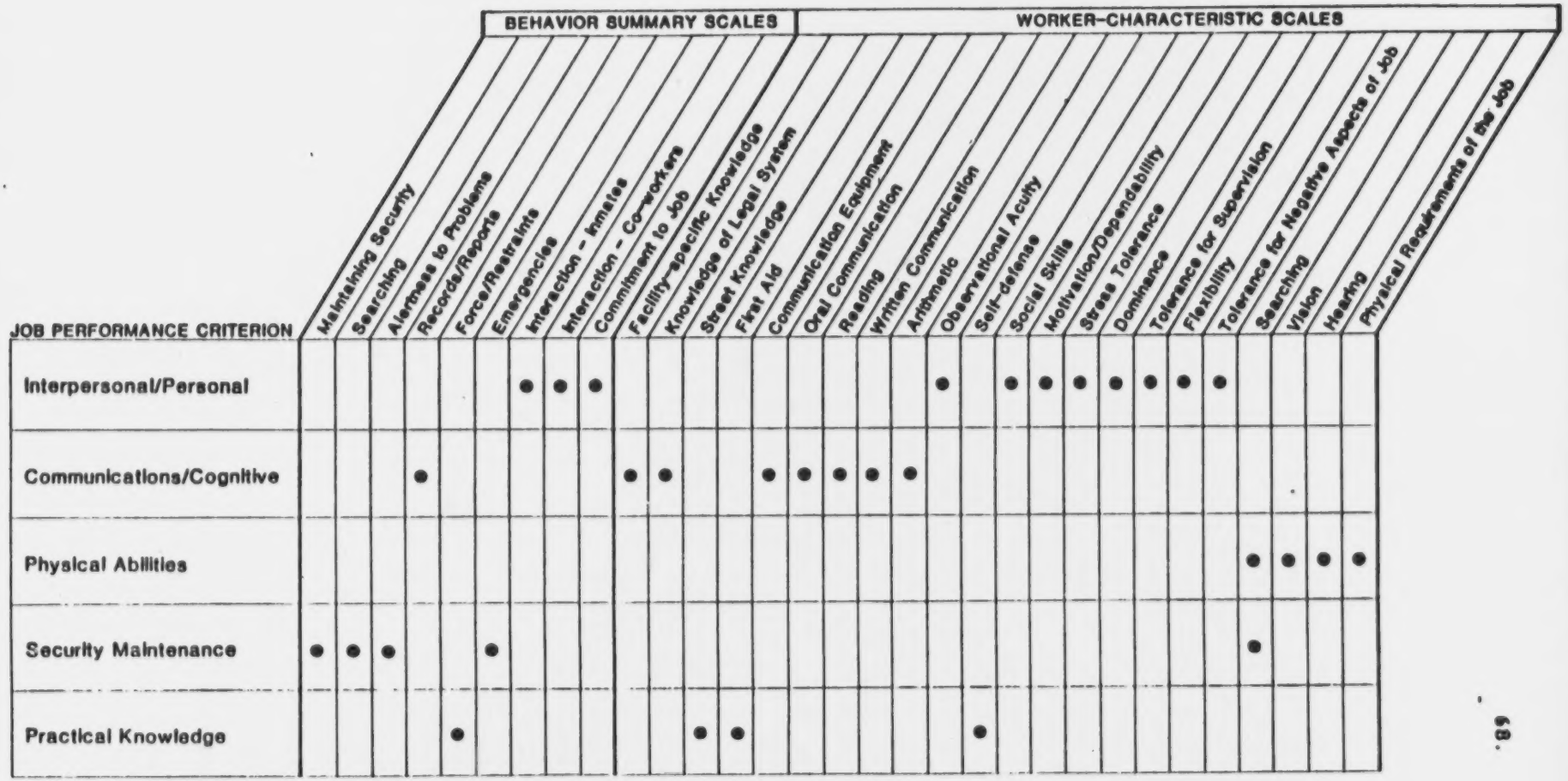
Criterion scores were determined by averaging across the standard scores (T-scores) of the rating scales loading on a particular factor. This method of computing factor scores differs from more commonly used methods in that in the latter, factor loadings are used to weight the contribution of each variable to each factor score. In such an approach every variable contributes to a factor score--even those

Table 6.3

Factor Descriptions of
Five Specific Job Performance Criteria

Criterion	Description
Interpersonal/Personal Skills	This criterion involves skills necessary for harmonious relationships with others. These skills involve flexibility, tolerance, and motivation, as well as skills in interacting with inmates and counselors.
Communications/Cognitive Skills	This criterion revolves around the 3 R's -- reading, writing, and arithmetic. It involves the ability to speak and write well, and requires knowledge of the facility one works in as well as general knowledge about the legal system.
Physical Abilities	This criterion requires good vision and hearing as well as skill in searching and the ability to meet the physical requirements of the job.
Security Maintenance	This criterion assesses officers' ability to remain alert to potential problems, to deal with emergencies, to keep accurate counts and records, and to be skilled in conducting searches.
Practical Knowledge	This criterion is related to an officer's knowledge of the rules and procedures appropriate to a given circumstance. Practical skills such as first aid techniques, how to use force and restraint (as opposed to the strength to do so) are most important here.

Figure 6.1 Simple Structure Factor Loadings of 31 Ratings Scales on Five Specific Job Performance Criteria



● Indicates that the rating scale was used to define the job performance criterion and contributes to an incumbent's score for that criterion.

variables not considered to be part of a particular factor. The approach used in this project was to assign a rating scale only to the factor on which it loaded most clearly. Factor scores were the unit weighted average of all ratings that define a given factor. Due to the nature of factor analysis, these factor scores, although derived from an orthogonal (varimax) rotation, are not uncorrelated.

To summarize, the difference between typical methods of computing factor scores and the method used in this project is that the typical method yields factor scores that are uncorrelated but that are computed using variables not included in the factor definition. The current method yields factor scores based only upon those variables that define each factor, but which are substantially intercorrelated. Table 6.4 presents the descriptive statistics for each of the five additional criteria as well as the intercorrelations of the five factor scores.

Criterion descriptive statistics. The criterion analyses yielded a total of six criteria: a criterion of overall effectiveness and five criteria of the more specific aspects of job performance (Interpersonal/Personal Skills, Communications/Cognitive Skills, Physical Abilities, Security Maintenance, and Practical Knowledge). Table 6.5 presents descriptive statistics for these criteria broken out by incumbent gender and ethnic status (the latter only for those groups represented in numbers sufficient to permit accurate calculations). For all six criteria, female corrections officers received ratings that were considerably below those received by male corrections officers. On the other hand, minority corrections officers were rated as performing better than nonminority officers on two of the criteria (Interpersonal/Personal Skills, Physical Abilities) and the same as nonminority officers for three criteria (Overall Performance, Practical Knowledge, and Security Maintenance). On only one criterion (Communications/Cognitive Skills) were minority corrections officers rated as performing at a level below that of nonminority officers. There was concern that the lower performance ratings of female corrections officers on all criteria and of minority corrections officers on the cognitive skills criterion might have been due to bias by the raters rather than due to true differences in the average performance. This concern led to the calculation of the average criterion scores by the ethnic status and gender of raters as well as by the ethnic status and gender of incumbents. Data for gender are given in Table 6.6. Care must be taken not to overinterpret these data as there are very few ratings available from female supervisors. What data there are, however, give no indication that the lower average criterion scores of female incumbents are attributable to bias by the supervisors. Bias would be indicated if male raters rated female incumbents lower than male incumbents, and if the ratings of female raters showed no difference between male and female incumbents. Table 6.6 indicates that supervisors of both sexes rated female incumbents lower than male incumbents. Female supervisors rated female incumbents lower than did male supervisors.

Table 6.4

Descriptive Statistics and Intercorrelations
for Five Factor Score Criteria

Criterion	Avg.	S.D.	Correlation with: 1	2	3	4	5
1. Interpersonal/Personal	50.3	7.8	--				
2. Communications/Cognitive	50.5	7.8	.72	--			
3. Physical Abilities	50.3	8.2	.71	.67	--		
4. Security Maintenance	50.4	8.1	.71	.70	.77	--	
5. Practical Knowledge	50.3	7.7	.69	.68	.62	.71	--

Table 6.5

Descriptive Statistics for the Overall and Specific Job Performance Criteria

Group		Overall	Specific Criteria[a] 1	2	3	4	5
Gender							
Males (192)	Mean	50.9	50.8	51.1	51.0	51.0	51.2
	SD	7.6	7.8	7.7	8.4	8.2	7.8
Females (58)	Mean	48.4	48.4	48.3	48.1	48.5	47.1
	SD	7.0	7.5	7.1	7.1	7.8	6.8
Ethnic Group							
Blacks (44)	Mean	50.2	51.2	47.9	49.3	49.4	49.4
	SD	7.4	7.4	8.4	7.5	9.1	7.2
Hispanics (55)	Mean	51.1	51.7	50.9	52.7	51.6	51.2
	SD	8.7	8.5	8.4	9.0	8.5	8.7
Whites (137)	Mean	50.3	49.8	51.2	49.7	50.2	50.4
	SD	7.1	7.3	6.9	8.1	7.7	7.6

[a] 1 = Interpersonal/Personal Skills

2 = Communications/Cognitive Skills

3 = Physical Abilities

4 = Security Maintenance

5 = Practical Knowledge

Table 6.6

Average Overall Criterion Scores by
Supervisor and Incumbent Gender

	Supervisor	
Incumbent	Male	Female
Male		
N	207	5
Mean	50.5	50.7
S.D.	7.7	1.9
Female		
N	55	9
Mean	48.8	46.8
S.D.	6.8	4.9

Ratings data for ethnic groups are given in Table 6.7. As was true for rater gender, care must be taken not to place too much emphasis on the sparse data in this table. The data indicate that there were slight, but statistically significant [$F(2,246) = 6.4$, $p < .01$] differences in the average ratings made by raters from different ethnic groups. Black and Hispanic supervisors gave higher ratings than did white supervisors, but there was no statistical interaction between supervisor and incumbent ethnic groups. That is, white supervisors were equally strict in their ratings of incumbents of all ethnic groups (in fact, rating black and Hispanic incumbents slightly higher than white incumbents). Taken together, the data in Tables 6.6 and 6.7 give no indication that the mean differences in criterion scores of subgroups are due to anything besides true differences in the ability of members of those groups to perform the corrections officer job tasks.

Predictor Analyses

Predictor analyses were conducted to ensure that all predictor data were accurately collected and scored, to evaluate the 15 tests in terms of their psychometric quality, and to describe the performance of the validation sample on each of the tests.

Data collection and cleaning. Incumbents recorded their test answers on machine readable answer sheets. Upon receipt of the answer sheets, the first step in predictor analysis was to check each sheet to ensure that all marks were dark enough for the optical scanner to "see" and to make sure that the ID number on every answer sheet corresponded to the number printed on the test booklet.

Once checked, the answer sheets were read by an optical mark reader and incumbents' responses were transferred to the computer where they would be further scored and analyzed. As part of scanning, any answer sheet that contained one or more apparently missing responses was flagged for inspection. This was necessary to ensure that marks too light for the scanner to "see" were not missed. Where necessary, marks were darkened and the answer sheets rescanned.

Scale revisions and final psychometrics. Once all predictor data had been accurately scanned, the 15 tests comprising the experimental selection battery were scored and examined for indications of needed revisions. The goal of scale revision was to achieve a final set of reliable tests for consideration during the validation analyses.

Scale revisions began with the computation of item-total correlations and item difficulty indices. Item-total correlations were computed using the point-biserial correlation and item difficulty was computed as the proportion of incumbents passing each item. Both statistics were computed for every item and distractor.

The item statistics were computed using data from all 339 completed experimental selection batteries. Items for which the keyed alternative correlated negatively with total scores were dropped from the tests. Additionally, typographic errors during the assembly of the experimental selection battery resulted in one item being dropped from each of the Reading, Arithmetic, Vocabulary, and Verbal Reasoning tests.

Item difficulty statistics for the Verbal Reasoning test indicated that few incumbents were successfully completing the last five items on the test and so the items were dropped.

Psychometrics for all tests (after revisions) are given in Table 6.8. Reliability coefficients (internal consistency KR-20) were computed and are presented for the cognitive ability tests; published estimates are given for the noncognitive tests. As can be seen, the tests are all sufficiently reliable to be used in pre-employment screening. In addition, the tests were all of moderate difficulty indicating that they were appropriate for use in the candidate population.

Descriptive statistics. Table 6.9 gives the mean (average score) and standard deviation of each test in the experimental selection battery. Table 6.10 shows the intercorrelations among those 15 tests. Tables 6.11 and 6.12 give the descriptive statistics of each test broken out by incumbents' gender and ethnic status, respectively. With the exception of Language Usage, female corrections officers performed more poorly on all of the cognitive ability tests than did their male counterparts. Similar but less pronounced differences between male and female corrections officers were found for the noncognitive tests. Females received higher scores for the Sociability (Sy) and Empathy tests and scored approximately the same as males on the Responsibility (Re) test. For the remaining noncognitive tests, however, male corrections officers received higher scores than did female corrections officers.

Black and Hispanic corrections officers received scores that were considerably below those of white corrections officers for all of the cognitive ability tests. For the noncognitive tests, minority corrections officers scored higher on the Sociability test than did nonminority corrections officers and below nonminority corrections officers on Empathy and COIB. For the remaining noncognitive tests the scores of minority and nonminority corrections officers were approximately the same.

Table 6.7

Average Overall Criterion Scores by
Supervisor and Incumbent Ethnic Status

	Supervisor		
Incumbent	Black	Hispanic	White
Black			
N	7	3	38
Mean	54.9	47.9	49.3
S.D.	5.2	13.5	6.6
Hispanic			
N	5	5	42
Mean	54.2	56.3	49.9
S.D.	6.7	12.3	8.5
White			
N	10	10	129
Mean	53.4	54.0	49.1
S.D.	7.2	5.3	7.2

Table 6.8

Final Item Statistics for the
Experimental Selection Battery

Test	Number of Items Original	Dropped	Avg. Disc.	Avg. Diff.	rel.[a]
Observational Memory	30	--	.15	.77	.56
Reading	70	24	.28	.80	.79
Arithmetic	35	1	.33	.76	.85
Language Usage	35	--	.26	.71	.76
Vocabulary	35	1	.33	.78	.83
Analogies	35	--	.30	.66	.81
Verbal Reasoning	30	6	.34	.61	.80
Sociability	35	2	.22	.68	.70-.85
Empathy	29	6	.12	.54	.84
Control	24	--	.34	.77	.84
Socialization	49	6	.19	.73	.65-.80
Responsibility	36	1	.19	.66	.65-.85
Stress Reaction	24	--	.48	.79	.90
Dominance	44	5	.26	.66	.65-.80
COIB	58	--	.10	.65	.70

[a]Reliability for noncognitive recognition tests are from published sources; the remainder are KR-20 estimates.

Table 6.9

Descriptive Statistics on 15 Individual Tests for the Final Validation Sample

Test	N[a]	Avg.	Stnd. Dev.
Observational Memory	335	23.14	3.23
Reading	339	37.02	5.22
Arithmetic	339	25.93	5.25
Language Usage	339	25.02	4.59
Vocabulary	339	26.61	5.00
Analogies	337	23.04	5.40
Verbal Reasoning	325	14.62	4.68
Sociability	337	22.54	4.39
Empathy	337	12.45	3.00
Control	336	18.37	4.11
Socialization	336	31.23	4.69
Responsibility	337	22.95	4.26
Stress Reaction	336	19.03	5.05
Dominance	336	25.91	5.52
COIB	325	29.95	3.95

[a]Ns less than 339 are due to tests not being scored for individuals with excessive missing responses.

TABLE 6.10

Intercorrelation of 15 Predictor Tests

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Obs. Memory	-														
2. Reading	.17*	-													
3. Arithmetic	.13*	.46*	-												
4. Language Usage	.13*	.51*	.51*	-											
5. Vocabulary	.22*	.59*	.46*	.60*	-										
6. Analogies	.21*	.54*	.52*	.51*	.56*	-									
7. Verbal Reas.	.22*	.55*	.43*	.38*	.44*	.57*	-								
8. Sociability	-.01	.12*	.11*	.08	.10*	.12*	.12*	-							
9. Empathy	.09	.20*	.15	.23	.19*	.26*	.20*	.52*	-						
10. Control	.04*	.09*	.05	-.03	.07	.01	.02	.08	-.08	-					
11. Socialization	.12*	.25*	.15	.12*	.11*	.18*	.13*	.32*	.20*	.42*	-				
12. Responsibility	.02	.22*	.21*	.23*	.26*	.23*	.19*	.45*	.34*	.32*	.45*	-			
13. Stress Reaction	.06	-.16*	.15*	.07	.11*	.13*	.10*	.48*	.32*	.38*	.47*	.44*	-		
14. Dominance	.07	.23*	.16*	.18*	.21*	.17*	.21*	.63*	.49*	.13*	.28*	.50*	.45*	-	
15. COIB	.15*	.20*	.15*	.12*	.09*	.15*	.18*	.12*	.14*	.24*	.49*	.27*	.30*	.16*	-

*Statistically reliable correlation at alpha = .05.

Table 6.11

Descriptive Statistics on Individual Tests
by Incumbent Gender

	Males[a]		Females[b]			
	Avg.	S.D.	Avg.	S.D.	t	ES[c]
Observational Memory	23.5	3.0	22.7	3.1	-1.69	-.24
Reading	37.6	4.2	35.8	6.9	-2.63[d]	-.37
Arithmetic	26.5	4.9	24.5	6.2	-2.71[d]	-.38
Language Usage	25.0	4.4	25.6	4.7	0.97	.14
Vocabulary	27.5	4.3	25.0	5.6	-3.85[d]	-.53
Analogies	23.6	4.9	22.7	6.3	-1.25	-.17
Verbal Reasoning	15.1	4.6	13.4	4.8	-2.58[d]	-.37
Sociability	22.5	4.5	23.2	4.5	1.04	-.15
Empathy	12.4	3.0	12.9	3.2	1.04	.14
Control	18.7	3.9	17.2	4.8	-2.63[d]	-.37
Socialization	31.5	4.6	31.0	4.2	-0.75	-.11
Responsibility	23.1	4.1	23.2	4.6	0.09	.01
Stress Reaction	19.5	4.7	18.5	5.5	-1.43	-.20
Dominance	26.1	5.6	25.7	5.4	-0.47	-.07
COIB	30.5	3.8	29.1	2.8	-2.69[d]	-.38

[a] N = approximately 214

[b] N = approximately 68

[c] ES = standardized mean difference, negative values indicate that females scored, on average, below males.

[d] statistically significant difference at alpha = .05

Table 6.12

Description Statistics on 15 Individual Tests
by Incumbent Ethnic Status

Test	Blacks[a] Avg.	S.D.	Hispanics[b] Avg.	S.D.	Whites[c] Avg.	S.D.	F	ES[d]
Obs. Memory	22.1	2.6	22.7	2.9	23.9	3.1	8.180[e]	-.49
Reading	36.4	5.6	34.9	6.7	38.3	3.7	11.467[e]	-.56
Arithmetic	25.2	5.6	24.3	6.2	26.9	4.6	6.400[e]	-.43
Language Usage	23.6	5.3	24.3	4.8	25.9	4.0	6.649[e]	-.44
Vocabulary	24.4	5.8	25.5	4.8	28.2	3.9	16.530[e]	-.70
Analogies	21.4	5.6	21.2	5.3	24.8	4.6	15.795[e]	-.70
Verbal Reasoning	13.5	4.6	12.3	4.5	15.9	4.4	15.040[e]	-.68
Sociability	23.7	3.7	22.4	4.8	22.4	4.6	1.530	-.12
Empathy	11.7	3.1	11.8	2.9	13.0	3.0	4.973[e]	-.39
Control	17.9	4.9	18.6	3.8	18.5	4.1	0.456	-.04
Socialization	30.7	4.9	31.7	4.4	31.6	4.5	0.752	-.07
Responsibility	23.1	4.3	22.8	5.0	23.3	3.9	0.264	-.08
Stress Reaction	18.9	5.0	20.0	3.8	19.0	5.1	1.025	.11
Dominance	26.2	5.1	26.0	5.7	25.9	5.6	0.064	.04
COIB	28.8	3.1	30.2	3.4	30.4	3.8	3.610[e]	-.25

[a]N = approximately 50

[b]N = approximately 60

[c]N = approximately 157

[d]Standardized mean difference contrasting blacks and Hispanics to whites; negative value indicates lower average scores by blacks and Hispanics.

[e]Statistically significant at alpha = .05

Validity and Fairness Analysis

Validity and fairness analyses had as their basic goal the evaluation of the potential of each test to provide valid, fair predictions of corrections officer job performance. Valid tests are those for which high scores indicate a higher probability of successful job performance than do low scores. Fair tests are those for which no average score differences occur among gender or ethnic groups except where differences of a similar magnitude occur for the criterion.

Once the validity and fairness of each individual test was determined, a combination of the tests could be chosen to form the Corrections Officer Examination, thereby meeting the project goal of developing and validating entry-level selection procedures for the corrections officer position.

An additional goal of the validity and fairness analyses was to determine whether any of the five specific job performance criteria were differentially predictable in a fair and valid manner, either by individual tests or by a different combination of tests than that chosen to predict overall job performance.

Adverse impact and test fairness. Tests for which two or more groups receive scores that are statistically different on average are said to discriminate between the groups. Used in this limited sense, the term discrimination simply indicates that there is a statistically reliable difference between the groups (i.e., a difference likely to be found every time the groups are tested). The presence of group differences for a selection procedure does not automatically disqualify that procedure however. The Uniform Guidelines on Employee Selection (1978) discuss the relationship between two important concepts related to the fair use of selection procedures. The first of these is adverse impact and the second is test fairness.

Adverse impact (as defined by the 4/5ths rule of thumb) exists whenever the selection ratio for members of a protected group is 4/5ths (80%) or less than that of the group with the largest selection ratio. Unfairness is defined in the guidelines as occurring when members of one group obtain lower scores on a selection procedure and when similar differences are not found for measures of job performance. The Guidelines also note that, if technically feasible, test fairness must be investigated whenever a selection procedure results in adverse impact, but not when adverse impact is absent.

From the perspective of selection procedures, discrimination in the limited sense described above can be either fair and therefore permissible, or unfair and therefore not permissible. In evaluating a selection procedure the first step is to determine whether adverse impact exists. Adverse impact is a joint function of the magnitude of group differences and the overall selection ratio used during selection. Even very large group differences may not result in adverse impact if a high enough selection ratio is set.

Many different models have been offered for use in distinguishing fair from unfair discrimination. One model (Cleary, 1968) holds that group differences in predictor scores can only be considered fair if differences of the same magnitude occur for the criterion measure. In statistical terms, this translates into a test of the similarity of regression equations computed separately for each subgroup involved in the fairness analysis. Selection decisions for any valid test or procedure depend upon the use of a single regression equation to predict the expected performance of all candidates. A selection procedure is unfair to the members of any group whose performance is systematically underpredicted by a common regression equation.

The standardized mean difference between groups (effect size) is an important statistic in this determination of test fairness. In the Cleary model, tests biased against minority groups (large negative effect size) will only be fair if a difference of comparable magnitude exists for criterion scores. From the perspective of adverse impact, selection procedures with large negative effect sizes require that very low passing scores be set in order that sufficient numbers of protected group members be selected. Effect sizes indicating bias towards protected groups warn of the necessity of further analyses to determine whether adverse impact is present, and, if adverse impact exists, whether the test is fair to protected groups.

Validity and fairness of the individual tests. Following all necessary revisions, each of the tests in the experimental selection battery was correlated with the overall criterion. At the same time, average scores and associated descriptive statistics were computed for all members of the final validation sample as well as separately by incumbent gender and ethnic group. These data are reported in Tables 6.11 - 6.13. As can be seen in Table 6.13, for five tests (Reading, Arithmetic, Language Usage, Vocabulary, and COIB) the validity estimates were statistically significant at the .05 level. Of these five, Reading stands out as the best single predictor of job performance when magnitude of the validity coefficient is the sole criterion of "best predictor". The validity coefficients of three additional tests (Analogies, Verbal Reasoning, and Control) were marginally significant, but none of the other tests demonstrated any promise of providing accurate predictions of corrections officer job performance.

Tables 6.11 and 6.12 presented descriptive statistics by gender (Table 6.11) and ethnic status (Table 6.12) as well as fairness statistics for

Table 6.13

Correlations of 15 Individual Tests
with Overall Job Performance

Test	Corr. w/job perf.	p[a]
Observational Memory	.06	.18
Reading	.24	.00
Arithmetic	.14	.02
Language Usage	.11	.04
Vocabulary	.14	.01
Analogies	.10	.06
Verbal Reasoning	.09	.07
Sociability	-.04	.26
Empathy	-.06	.19
Control	.09	.08
Socialization	.01	.44
Responsibility	.06	.19
Stress Reaction	-.00	.47
Dominance	.08	.10
COIB	.14	.02

[a]Probability that a correlation this large or larger would be obtained due to chance. Probabilities below .05 are considered statistically significant.

each test. For each, means and standard deviations by subgroup are given, along with the appropriate test for determining whether differences between groups are statistically significant. Finally, an index of the magnitude of the difference between groups is reported. This index, effect size, gauges the distance between group means relative to the pooled score variance. Conceptually, effect size estimates are interpreted in the same way as are z scores. For example, an effect size of .30 indicates that the mean of one group lies .3 pooled standard deviations above the mean of the second group and, therefore, that 62% of the members of the first group receive scores greater than the average score of the second group. In groups with equal means, 50% of persons in either group receive scores above the mean of the other group.

For gender, the data in Table 6.11 show that all of the tests with statistically significant validity coefficients (except Language Usage) also had statistically reliable mean differences favoring males: Language Usage scores favored females, but the difference between average male and female scores was not statistically reliable. Even though scores on these tests statistically favored males, the average differences between males and females are (for all tests except Vocabulary) of equal or lesser magnitude than the average difference in criterion performance. Thus with the single exception of the Vocabulary test, those tests shown to be valid predictors of job performance are also fair predictors for males and females. Because the effect size for the Vocabulary Test exceeds the effect size for the criterion, Vocabulary, when considered alone, is a valid but unfair predictor of corrections officer job performance.

Fairness data for the three ethnic groups with large enough samples to permit calculation of reliable statistics are given in Table 6.12. The statistics in Table 6.12 are quite similar to those reported for gender except that a oneway analysis of variance rather than a t test was used to determine whether the differences among groups were statistically significant. The effect size measures for ethnic groups contrast combined black and Hispanic officers to white officers.

The findings for ethnic groups are similar to those for gender. All tests (except Control) for which statistically significant validity coefficients were found also displayed statistically reliable differences among ethnic groups. In all cases these differences favored whites. The findings for ethnic status differ from those for gender however, in that the magnitude of the mean differences among ethnic groups, while of comparable size to that between males and females, is considerably greater than the corresponding difference in criterion performance. Thus, for ethnic groups, the tests showing promise as valid measures of job performance are not equally fair predictors for the three ethnic groups.

Developing a final selection battery. Following the evaluation of the fairness and validity of each of the 15 individual tests, it was necessary to determine which combination of those tests would be recommended for use as the Corrections Officer Examination. Several constraints influenced the choice of tests to be included in the final battery. Principal among these was the necessity that the battery be valid and either, incur no adverse impact, or provide fair predictions of the performance of candidates from all groups. Of considerable additional importance was that the final selection procedures assess candidates' potential in all important job areas, that the final set of tests require no more than 2 1/2 hours to administer, and that scoring be as simple as possible so as not to compromise the operational validity of the final battery.

A core set of tests was selected from those that made up the experimental selection battery by selecting the two noncognitive tests which, by themselves, were at least marginally valid predictors of the overall criterion (COIB and Control). These tests were then augmented with the two most fair and valid cognitive tests (Reading and Language Usage). A total score was computed for each corrections officer incumbent by summing his/her scores on each of the four selected tests. This combination of tests was found to be a valid predictor (r with overall performance = .21, $p < .05$). It resulted in a sizeable mean difference favoring whites (effect size = -.50), suggesting that adverse impact would be present.

Several procedures are available that permit the use of apparently unfair tests in a manner that does not result in adverse impact against protected groups. For example, separate selection lists can be constructed and candidates selected from both lists in numbers that do not violate the 4/5ths rule of thumb. Another method is to adjust test scores by adding a constant value to some scores, thus removing the mean differences among groups. As neither of these methods were administratively feasible for the Board of Corrections, the following method was used to expand the final selection battery while maintaining or increasing its validity and decreasing the magnitude of the mean differences among ethnic groups.

Because none of the remaining tests from the experimental selection battery could be used without increasing the bias against minority groups, a search was undertaken at the item level. Because the cognitive ability tests were highly intercorrelated and because they were the principal source of mean differences favoring whites, the pool of items to be searched was narrowed to those items from the noncognitive tests that were not already in the final test battery.

The final validation sample was randomly split into halves and the effect size of each item (blacks and Hispanics versus whites) was computed for every item in both samples. Items with positive item validities and that favored minorities in both samples were selected for use in the final battery.

In reviewing these statistics, it was noted that several items from two noncognitive tests (Sociability and Empathy) would favor minorities in both random halves if the keying of those items was reversed. The content of each of these items was examined to ensure that reversing their keys would not result in rewarding candidates for endorsing undesirable traits or behaviors. As long as reversing an item key was judged not to cause such problems, the keys were reversed and the items added to those already selected for the battery.

This procedure identified 38 items drawn from a maximum possible 197 items from the six noncognitive tests not included in the final battery. Table 6.14 indicates the source of the items selected using the methods described above. Predictor scores were recomputed for all incumbents by summing across the items that comprise the Corrections Officer Examination. These scores were found to correlate .24 with the overall criterion, indicating that the battery was a valid measure of job performance.

Although every precaution was taken to minimize its occurrence, the method used to select additional items had some probability of having capitalized on chance variation among items. The effect of such capitalization would be to slightly overestimate the validity of the scores. Thus, the true validity lies somewhere in the interval from .21 (validity prior to the addition of the 38 items) to .24 (validity after the addition of the 38 items).

Fairness data for the final battery are given in Table 6.15. As can be seen, the mean difference between male and female corrections officer scores is statistically significant with an effect size of -.37 compared to an effect size of -.34 for the criterion (a difference that is itself also statistically significant). For ethnic groups the mean differences among predictor scores were not statistically significant and their magnitude was -.13 as contrasted with a nonsignificant difference with effect size .05 for the overall criterion. That the mean differences among ethnic groups are not statistically significant indicates that they are no larger than would be expected due to chance differences that occur between any groups of similar size. Nevertheless, the predictor effect size for ethnic groups was larger than the corresponding criterion effect size indicating the need to examine whether the Corrections Officer Examination resulted in adverse impact against minority candidates. Additionally, the predictor effect size for gender was large enough that use of the predictor may have resulted in adverse impact against female candidates. As a result, the adverse impact of the final predictor was examined for female candidates as well.

Table 6.14

Source of Items Selected to
Augment the Final Selection Battery

Source	Items Selected[a]
Sociability[b]	6
Empathy[b]	8
Responsibility	8
Socialization	6
Dominance	10
Stress Reaction	2

[a]Number selected sums to 40 due to two items that are scored on two scales.

[b]Items are reverse keyed.

Table 6.15

Gender and Ethnic Group Statistics
for Corrections Officer Examination Scores

Group	N	Mean	S.D.	F	ES
Males	214	136.1	11.3		
Females	68	131.5	15.2	6.96[a]	-.37
Blacks	50	133.7	11.9		
Hispanics	60	133.2	17.8		
Whites	157	136.2	10.2	0.57	-.13

[a]Statistically significant at alpha = .05

As discussed previously, effect size is one of two critical factors in determining whether a test will violate the 4/5ths rule of thumb; the other factor being the cutoff score chosen. Scores on the Corrections Officer Examination were examined to determine which cutoff scores, if any, would result in a violation of the 4/5ths rule, and thus have adverse impact against female or minority corrections officers.

These analyses showed that, for minorities, as long as the overall selection ratio was greater than .10 (raw score of 149), no violation of the 4/5ths rule occurs. A selection ratio of .10 is much more stringent (only 10% of all applicant pass) than is appropriate for a battery of this type. Even if such a rigorous cutting score were set, use of the final selection battery to "select" candidates from the research sample would not have resulted in adverse impact against minority group members. Assuming that the effect size for ethnic groups remains near -.13 in applicant groups and further assuming that minority and nonminority groups have approximately equal score distributions, in order to avoid adverse impact, selection ratios must be kept above .20, a value above that recommended for use with this (see Chapter 7). As no adverse impact was found for minority corrections officers, no further analyses were conducted.

Adverse impact analyses for female officers revealed that the Corrections Officer Examination would result in adverse impact for almost every reasonable selection ratio (i.e., all selection ratios .80 or less). As per the Uniform Guidelines, the finding of adverse impact was followed by investigation into the fairness of the Corrections Officer Examination for females. These analyses consisted of computing three regression equations: a common equation used to predict the criterion scores of both males and females, and a separate equation for males and a third for females. The general linear model was used to contrast the male and female regression equations. This contrast revealed no statistically significant differences [$F(2,235) = 1.6$, $p > .05$] between the equations. The finding of no statistically significant difference in regression equations indicates that the job performance of females would not be underpredicted by the use of a common prediction equation. This finding, coupled with the similarity in effect sizes for predictor and criterion scores, indicates that the lower test scores of female corrections officers in this project are accurate indications of their somewhat lesser ability, as a group, to perform the corrections officer job. Thus, the Corrections Officer Examination, although yielding mean differences that favor males, is a valid and fair predictor for both men and women.

The Uniform Guidelines also indicate that when a selection procedure results in adverse impact, it is incumbent upon employers to consider alternative procedures that would result in less adverse impact. There was no combination of the tests included in the experimental selection battery that resulted in less adverse impact against females without either: a) introducing adverse impact against minority candidates; or b) leading to an invalid procedure.

Differential predictability of specific criteria. Correlations of the 15 tests and the final predictor with five specific criteria of job performance are shown in Table 6.16. As can be seen, there were several individual tests whose use resulted in valid predictions of one or more of the specific criteria. As can also be seen, the final predictor was as valid, or more valid a predictor than were any of the individual tests. Tables 6.11 and 6.12 report the likely adverse impact of the individual tests against female and minority candidates. Because the individual tests invariably result in considerable adverse impact, none of them would be both fair and more valid than would the Corrections Officer Examination. It was therefore concluded that the specific criteria were most predictable using the Corrections Officer Examination scores, so no additional predictors were investigated.

Table 6.16

Correlations of Individual and Final
Predictors with the Five Factor Criterion Measures

Predictor	Factor Criterion[b] 1	2	3	4	5
Observational Memory	02	09	05	08	07
Reading	20[a]	25[a]	07	16[a]	21[a]
Arithmetic	10	20[a]	16[a]	09	09
Language Usage	07	18[a]	10[a]	08	08
Vocabulary	14[a]	27[a]	10[a]	09	12[a]
Analogies	03	19[a]	03	09	08
Verbal Reasoning	04	15[a]	01	09	04
Sociability	-01	-09	-08	-07	-03
Empathy	-06	-02	-09	-02	-02
Control	11[a]	-02	-01	05	09
Socialization	04	-03	-05	-03	08
Responsibility	08	02	-02	02	03
Stress Reaction	02	-05	02	01	07
Dominance	08	11[a]	02	05	13[a]
COIB	13[a]	10[a]	09	07	16[a]
Final Predictor	26[a]	23[a]	13[a]	16[a]	26[a]

[a]Statistically significant at alpha = .05
decimal points omitted

[b]1 = Interpersonal/Personal Skills
2 = Communications/Cognitive Skills
3 = Physical Abilities
4 = Security Maintenance
5 = Practical Knowledge

CHAPTER SEVEN: DISCUSSION AND RECOMMENDATIONS

The Corrections Officer Examination is a fair and valid indicator of potential to successfully perform the corrections officer job. PDI therefore recommends that the Board of Corrections adopt this selection instrument or some other equally valid and fair measure of the same abilities and personal characteristics) as part of the entry-level selection procedures for agencies participating in the STC program. The Corrections Officer Examination is not intended to serve as the sole or even as the primary basis for selection decisions. Its judicious use, in combination with other sources of information about candidates' potential (e.g., application appraisals, interviews, educational achievement, etc.). and according to the constraints set forth below will, however, increase the probability that selected candidates will prove to be successful corrections officers.

The remainder of this chapter provides a description of the format and content of the Corrections Officer Examination and sets forth the more important constraints governing the use of the examination as one source of information about candidates' potential. Recommendations are made concerning: a) scoring of the examination; b) the data collection and reporting necessary to supplement the notes collected in this project; and c) the efforts necessary to document the continued validity and fairness of the examination.

Exam Format and Content

The Corrections Officer Examination consists of a single test booklet containing items assessing candidates' cognitive and noncognitive knowledges, skills, abilities, and personal characteristics. The examination contains 195 items divided into four parts and has a total testing time, including time for instructions, of approximately 1 1/2 hours.

Two related cognitive abilities are measured in Parts 1 and 2. The first of these (Reading) requires that candidates read four brief (1-3 paragraph) passages and answer 10 to 12 questions about each passage. The questions are mostly factual and require only the ability to remember what was read (or to quickly review information). A few of the questions require that simple conclusions be drawn based upon information presented in each passage. Taken together, the Reading items measure candidates' ability to read, understand, and recall written information. The passages are written at approximately a twelfth-grade level (Flesch index) and are intended to be similar in content to the types of information corrections officers read on the job.

Part 2 contains items that measure candidates' ability to recognize common errors of grammar, punctuation, or capitalization. These items consist of 35 sentences, each divided into five sections. In most of the sentences there is an error of grammar (e.g., "Them did the job well."), punctuation (e.g., "The boy in the red, coat."), or capitalization (e.g., "On friday, he left town."). Candidates are to indicate which section of each sentence contains the error or to indicate that the sentence contains no errors. These items require a basic understanding of correct language usage and of the conventions of written English.

Part 3 of the Corrections Officer Examination contains 18 items from the COIB. Each item consists of a list of three activities. Candidates are to choose the activity they would like most and the activity they would like least. The personal characteristics measured by these items are described below.

In Part 4 the Corrections Officer Examination contains 96 items that assess several personal characteristics found (in the job analysis) to be important components of successful corrections officer performance. These items were drawn from scales of the California Psychological Inventory, the Multidimensional Personality Questionnaire, and the Correctional Officers' Interest Blank (COIB). Following are the personal characteristics measured by the items. These descriptions borrow heavily from descriptions of the scales where the items originally appeared.

Approximately one-third of the items were drawn from scales (MPQ Control and CPI Socialization and Responsibility) that reflect a tendency to take a cautious, planful approach to life. Persons who receive high scores on these scales are described by others as being rule-respecting, self-controlled, responsible, and preferring to plan their activities rather than play things by ear. Another personal characteristic measured is dominance (CPI Do). Dominant persons are those who behave in an ascendant manner, who are self-confident, who can take the initiative and exercise leadership in a self-assured manner.

A few of the items were drawn from CPI Sociability and Empathy scales but have had their keys reversed (see Chapter 6). As scored in the Corrections Officer Examination, these items reflect persons who dislike being the center of attention, who are autonomous (do not have a high need to seek out others for company), who prefer that things go smoothly and without interruptions, and who act upon principles of right and wrong even if that conflicts with their popularity or social relations.

Finally, the Corrections Officer Examination contains the entire COIB which research has shown to be strongly related to the CPI So and Re scales (also represented in the exam) and which identifies individuals "with strong internal values and who can be counted on to behave in a dependable rule observing way." COIB items with a true/false format are contained in Part 3 of the exam. Part 4 contains the remaining COIB items.

The Corrections Officer Examination is intended to be a power rather than speed measure. The time limits suggested for the examination are such that most candidates will have ample opportunity to complete all items. The examination equally weights the cognitive and noncognitive items and is scored by counting the number of items that a candidate answered correctly. Table 7.1 describes the scales from which examination items were drawn.

Using The Corrections Officer Examination

The Corrections Officer Examination is intended to be used only as described in the User's Manual. This manual provides general information concerning the physical settings recommended for testing as well as verbatim proctoring instructions. A Candidate Information Booklet has been prepared that briefs candidates about the purpose and nature of the examination. No candidate should be tested without having received the Candidate Information Booklet in time to review the information it contains.

The Corrections Officer Examination is a fair and valid indicator of candidates' potential to perform the corrections officer job but only if it is administered and scored according to the guidelines in the User's Manual. The validity and fairness of any other use of the examination cannot be guaranteed. Persons wishing to use the examination in a manner other than that described in the User's Manual must document for themselves the validity of the examination for that use. All items in the Corrections Officer Examination are copyrighted and permission to use the examination in part or in whole must be obtained from the California Board of Corrections, Consulting Psychologists Press and the University of Minnesota Press.

Exam Security

Security of the test keys is of the utmost importance. If the security of the keys is breached, the validity of the entire examination is compromised. The governing principle of key security is: The fewer people to have access to the keys the better. The best possible arrangement would be to have all tests scored centrally at the Board of Corrections and to allow only one or two persons access to the test keys. If central test scoring is not possible, the next best arrangement would be to allow only one person from each department access to test keys. This person would be responsible to the Board of Corrections for the maintenance of test security.

Test keys should not be reproduced indiscriminately. All copies of the keys should be numbered and an accurate record kept of their location.

Security of test keys also depends upon the security of the examination booklets. Test booklets should be numbered and accurate records kept by the Board of Corrections as to which agencies have which booklets.

Table 7.1

Description of the Noncognitive Tests
Contributing Items to the Corrections Officer Examination

Test	Description
CPI Dominance	A subscale of the California Psychological Inventory that identifies individuals who behave in a dominant, ascendant manner and who, in interpersonal situations, can take the initiative and exercise leadership in a forceful, self-confident manner.
CPI Sociability	A subscale of the California Psychological Inventory that indicates a person's tendencies to be outgoing and participative and to seek out and enjoy social contact without being shallow or superficial.
CPI Responsibility	A subscale of the California Psychological Inventory measuring a person's tendency to believe that life is best governed by reason and to behave in a responsible, self-disciplined manner.
CPI Socialization	A subscale of the California Psychological Inventory measuring a person's tendency to be socialized, rule-respecting, prudent, and responsive to what others feel and think.
MPQ Stress Reaction	A subscale of the Multidimensional Personality Questionnaire that indicates persons who are able to put fears and worries out of their mind easily; quickly get over upsetting experiences; and who are not troubled by emotional turmoil or feelings of guilt.
MPQ Control	A subscale of the Multidimensional Personality Questionnaire indicating a person's tendency to be reflective, cautious, careful, rational, and sensible, and who likes to anticipate and plan activities and events.

Table 7.1 (continued)

Test	Description
COIB	The Correctional Officers' Interest Blank is a 40-item questionnaire that "tends to identify individuals with strong internal values and who can be counted on to behave in a dependable and rule-observing way."
Empathy	A series of items drawn from the California Psychological Inventory that identifies persons who are described as being "characterized by a patient and forebearing nature, by affiliative and socially ascendant tendencies and by liberal and humanistic political and religious attitudes."

Periodic (semi-annual) inventories should be made to ensure that all testing materials can be accounted for and remain secure. If examination materials are lost or destroyed, the Board of Corrections must be informed immediately, so that the appropriate action can be taken.

Scoring and Data Collection

Detailed instructions for scoring the Corrections Officer Examination are contained in a separate confidential report. Briefly, scoring is accomplished via a set of scoring templates. Candidates will record their responses on an answer sheet. To score the examination, the appropriate scoring template is placed over the answer sheet and the number of responses visible through the template are counted. Parts 1-4 will be scored separately and overall scores formed by summing the Part 1 - Part 4 subscores. Space will be provided on the answer sheets to record candidates' scores. Examination answer sheets are designed to permit easy transfer to electronic scoring systems.

Table 7.2 contains a frequency distribution of incumbents' examination scores. These data show the number and proportion of candidates who received various final scores. Also shown are the actual and expected percentile equivalents of the scores. The actual percentile equivalents are simply the proportion of incumbents who received a score at or below a given value. Expected percentile equivalents reflect the proportion of incumbents who would score at or below a given score based upon a normal distribution of test scores with the same mean and standard deviation as were obtained in the incumbent sample. Because these data are based on the scores of incumbents, they may slightly overestimate the proportions of applicants expected to obtain any given score.

The data in Table 7.2 are the norms that PDI recommends be used to guide test scoring and interpretation for initial administrations of the Corrections Officer Examination. It is important that agencies record (at a minimum) the race, sex, and final score of every applicant tested with the Corrections Officer Examination. These data are essential for evaluating the appropriateness of the recommended minimum acceptable score and for monitoring the examination's validity and fairness. PDI recommends that the Board of Corrections also retain copies of all answer sheets so that item level analyses of the Corrections Officer Examination can be undertaken as necessary. If examinations are scored by individual agencies (not recommended), copies of all testing materials should be sent to the Board of Corrections semi-annually. Agencies should retain copies of all testing materials for a minimum of one year, as these materials are likely to be subpoenaed should the examination be challenged in court.

The Board of Corrections is strongly encouraged to conduct annual audits of the Corrections Officer Examination to, at a minimum, check scoring accuracy, evaluate the current recommended minimum acceptable score, and

Table 7.2

Corrections Officer Examination Score Distribution

Raw Score	N	Percentile Equivalent	
		Actual	Expected
164	295	99	99
155	287	97	95
151	278	94	90
148	259	88	85
146	247	84	80
144	229	78	75
142	210	71	70
140	191	65	65
139	180	61	60
137	159	54	55
135	137	46	50
134	122	41	45
132	100	34	40
131	93	32	35
129	79	27	30
127	67	23	25
125	54	18	20
123	43	15	15
120	28	10	10
115	18	6	5
107	9	3	1

evaluate gender and ethnic differences in test scores. It is also recommended that the examination be revalidated against a job performance criterion after it has been operational for two years and every three to five years thereafter unless otherwise indicated by the annual audits.

Score Interpretation and Cutoff Scores

Methods used to set cutoff scores and a recommended minimum acceptable score are contained in a separate report available from the Board of Corrections. Briefly, this report determines the probability that candidates scoring within several score ranges will be average or better performers. The minimum recommended acceptable score is that for which there is a 50-50 chance of being an average or better performer.

Although a minimum acceptable score has been recommended, no optimal, statewide cutoff was set. Choosing an appropriate cutoff score requires consideration of factors that will vary from agency to agency. This makes the establishment of a uniform statewide cutoff impossible. The cutoff score report discusses the factors that influence the setting of cutoff scores and provides guidelines for agencies to follow in establishing their own cutoff scores. This will allow individual agencies considerable flexibility in determining how best to use the Corrections Officer Examination.

CHAPTER EIGHT: TYPE I FACILITIES SURVEY

Corrections Officer Selection and Training

The original research project included a job analysis of corrections officers employed in Type I,II, III and IV facilities operated by county sheriffs' departments. This research identified the core tasks performed by 70% o the personnel in these facilities statewide. The entry-level selection procedures and training curricula were derived from the knowledges, skills, abilities and other personal characteristics necessary to perform these tasks.

Subsequent to the research project, subject matter experts from city operated Type I facilities and/or departments operating several Type I facilities reviewed these core tasks. the tasks that the experts identified as being performed by corrections officers in their Type I facilities were statistically cross-referenced to the research knowledge/skill maps and training curricula. The tasks identified as not being performed were statistically discounted.

Through this process it was determined that a few of the minor tasks were not performed. The remaining tasks still necessitated the measurement of the knowledge, skills, abilities and other personal characteristics identified through the research project and measurable by the entry-level selection procedures.

It was also determined that the tasks which could be discounted would have an insignificant impact on the training curriculum subject matter and instructional hours. The removal of 15 to 20 minutes of instructional time from a few modules of instruction "sprinkled" throughout three weeks of corrections officer training, would reduce the quality of the presentation and the consistency of the subject. These facts could not justify a curriculum different from other corrections officers in the state.

n view of the above, the decision was made that entry-level staff assigned primarily to Type I facilities would comply with the same selection and training standards as other corrections officers who are assigned to Type I, II, III and IV facilities in the state.

References

American Psychological Association. (1974). Standards for Educational and Psychological Tests. Washington, DC: Author.

American Psychological Association, Division of Industrial-Organizational Psychology. (1980). Principles for the Validation and Use of Personnel Selection Procedures. (2nd ed.). Berkeley, CA: Author.

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Equal Employment Opportunity Commission. (1978). Uniform Guidelines on Employee Selection Procedures. Federal Register, 43, 38290-38315.

Gough, H.B. (1982). Manual for the Correctional Officers' Interest Blank. Palo Alto, CA: Consulting Psychologists Press.

Overall, J.E., & Klett, C.J. (1971). Applied Multivariate Analysis (Ch. 7). New York: McGraw-Hill.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 86, 420-428.

STANDARDS AND TRAINING
FOR CORRECTIONS PROGRAM

STC

BOARD OF CORRECTIONS

TECHNICAL ADEQUACY REPORT

CORRECTIONS OFFICER

Copyright 1987 California Board of Corrections

In preparing the Corrections Officer Examination for the California Board of Corrections, Personnel Decisions, Inc. has taken into account the following documents concerning selection testing:

- The American Psychological Association, Standards for Educational and Psychological Tests, (1974).
- The APA Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures, 2nd edition, (1980).
- The Equal Employment Opportunity Commission, Uniform Guidelines on Employee Selection Procedures, (1978).

Report On the Technical Adequacy Of the Final Validation Sample For Corrections Officers

The original project plan for the PDI-STC Corrections Officer Validation Project called for testing a minimum of 500 corrections officer incumbents. For reasons elaborated herein, the actual number of officers in the final validation sample ranged from 250 to 339 (depending upon the analysis being conducted). This report documents the data collection efforts resulting in the final validation sample in order to show: (1) that every effort was made to collect usable data from as many incumbents as could reasonably be tested; and (2) that the number of incumbents actually tested is technically adequate to permit the reliable computation of all necessary statistics and analyses.

The data collected in this criterion-related validation project consisted of scores on 15 measures of cognitive and noncognitive skills and abilities (predictors), and job performance ratings for 31 specific areas, plus a rating of overall effectiveness (criteria). For most analyses (e.g., validity, fairness) it was necessary that data for both predictor and criteria be available while other analyses (e.g., norms) required only that predictor data be available..

Predictor Data Collection

Data collection began with telephone canvassing of all agencies that had participated in the job analysis phase of the project. Agencies were contacted to determine their willingness to participate in the testing phase of the project and to get an approximate census of the number of eligible incumbents from each agency (i.e., those on the job for six or more months). Target samples were based upon relative agency size, but smaller agencies were generally requested to test a proportionally larger number of their incumbents than were larger agencies. The initial targeted sample called for testing 497 corrections officers from 16 sheriff's departments.

Once targeted samples were set, participating departments were contacted by phone to reconfirm their participation in the testing phase of the project and to secure permission to test the targeted number of incumbents. During the follow-up phone calls, Los Angeles County informed us that, due to delays in the scheduled testing date (which was then in conflict with their preparation for the Summer Olympics), they would be unable to participate in testing. The withdrawal of the Los Angeles County Sheriff's Department reduced the overall targeted sample from 497 to 251.

The remaining 15 agencies were contacted again to see if additional incumbents were available for testing. The agencies committed an additional 39 incumbents to the testing phase, thus raising the total targeted sample to 290.

Because the total targeted sample was still well short of the 500 incumbents called for in the project plan, agencies that had participated in the job analysis phase of the project but were not participating during testing were contacted to see if they might be able to supply additional incumbents. Three agencies agreed to supply an additional 45 incumbents, raising the total targeted sample to 335.

To further raise the number of incumbents tested, STC consultants contacted a number of agencies that had not previously participated in any phase of the project. Due to the efforts of STC, three agencies joined the validation phase of the project and supplied an additional 55 incumbents, raising the total targeted sample to 390.

When it appeared that no additional incumbents were available for testing, materials were distributed to all agencies which had not received them and testing commenced. Unfortunately, not all agencies were able to meet their targeted sample and only 75% of the total targeted sample (289 of 390) was actually tested. Inspection of these data indicated that too few minority group incumbents had been tested to permit reliable calculation of necessary fairness statistics. STC consultants then arranged with the Los Angeles County Sheriff's Department to test 50 additional black or Hispanic corrections officers, bringing the total number of corrections officers tested to 339.

Criterion Data Collection

The initial sampling plan called for the collection of job performance ratings for all corrections officers tested. Ratings materials were supplied to participating agencies in sufficient quantity to permit rating of all tested incumbents. Coordinators were instructed to select incumbents for testing in a manner designed to minimize the administrative burden of collecting the job performance ratings. Nevertheless, and despite repeated follow-up contacts with agencies, ratings data were collected for only 282 of the 339 corrections officers for whom predictor data were also available.

Due to test booklets being returned without incumbents' names, incumbents with fewer than six months experience being tested, and ratings made by supervisors who indicated that they did not know the rated incumbents "well at all", the final number of corrections officers for whom both predictor and criterion data were available was 253.

Technical Adequacy of the Sample

There are three analyses in particular for which the technical adequacy of the final validation sample must be shown. These are the validation analysis, fairness analysis and the development of test norms. Each analysis involves one or more specific statistical

tests. Every statistical test involves the testing of a null hypothesis (e.g., no validity, or no difference between groups) versus some, usually unspecified, alternative hypothesis. If an estimate of the magnitude of the effect being tested can be specified, (e.g., validity greater than .30, mean difference of .80 standard deviations), the power of the test to detect that effect can be determined. Statistical power is simply the probability (expressed as the number of chances in 100) that the test will reject the null hypothesis if the effect being tested is as large or larger than its estimated value.

The power of all statistical tests is a direct function of the amount of data used to compute the test statistic -- the more data, the more powerful the test. A given sample is technically adequate for a given statistical test if it is of sufficient size to permit a reasonably powerful test of the expected effect. To determine the power of a statistical test, some estimate of the expected magnitude of the effect must be made. The smaller the expected effect, the larger the sample needed to achieve a reasonably powerful statistical test. Statistical power is directly analogous to a microscope where more power is needed to see very small objects than is needed to see relatively larger objects. The technical adequacy of the final validation sample is estimated in the following for the three main analyses of the project.

Validity analysis. The technical adequacy of the final validation sample for validity analyses depends upon the power of the statistical test of the significance of the correlation between predictor and criterion. The correlation coefficient is based upon data from those corrections officers for whom both predictor and criterion data are available. In order to determine the power of the statistical test for a sample of 253, some estimate of the expected validity is required. Boehm (1982) reports that the average validity coefficient from 176 studies spanning 20 years was .22. This is the estimate used to estimate the power of the validity analyses for this project.

Schmidt, Hunter, and Urry (1976) provide a formula to determine the sample size necessary to achieve a desired power for a specified validity coefficient. The problem here is only somewhat different than that posed by Schmidt et al. We want to know the power for a sample of 253 and an estimated validity of .22. The Schmidt et al formula can be converted for use in answering this question. Thus it can be shown (see Appendix) that if the true validity of the final predictor is at least .22, then 94 times out of 100 it will be detected as being statistically significant using a sample of 253. Thus it is clear that the final validation sample in this project is technically adequate for the purpose of validation analyses.

Fairness Analyses. The technical adequacy of the statistical tests involved in fairness analyses depends upon their ability to detect subgroup differences in average predictor and average criterion

scores. As was true for validity analyses, some estimate of the magnitude of the expected effect is necessary to determine the power of the statistical tests. While a straightforward estimate was available for validity, no such easily obtainable estimate exists for gender or race differences.

The final predictor in this study consists of a combination of cognitive and noncognitive measures that has not been previously examined. For general measures of intelligence, it is not uncommon to see differences of approximately one standard deviation between minority and non-minority groups, while males and females typically do not differ. For the cognitive component of the Corrections Officer Examination we might expect to find rather large ethnic group differences favoring whites but no gender differences. The magnitude of typical group differences for the tests that make up the noncognitive portion of the examination is not known. In any case, no sufficiently precise estimate of the expected magnitude of group differences on the Corrections Officer Examination is possible.

Without a reasonable estimate of the expected magnitude of group differences, no precise determination of statistical power can be made. Figures 1 and 2 show the power of the t test (a statistical test for group differences) as a function of the standardized mean difference between groups. Figure 1 reports the power of the test to detect male-female differences and Figure 2 shows power for detecting ethnic group differences. In Figure 2, three power curves are shown. These are, from left to right, power to detect differences between grouped minority and non-minority incumbents, Hispanic and white incumbents, and black and white incumbents. Note: Although not shown as such, the power curves in Figure 2 are symmetric about the Y-axis.

The power curves shown in Figures 1 and 2 indicate that, for the sample sizes in this project, the statistical test for detecting group differences has sufficient power to detect large group differences and reasonable power to detect moderate group differences (magnitude = .25 - .45 standard deviations), but that the power for detecting small group differences (0.1 - .25 standard deviations) is low.

Fairness analyses depend less upon detecting statistically significant differences between groups than they do upon estimating the magnitude of subgroup differences on predictor and criterion measures. For this reason, less powerful statistical tests can be tolerated for fairness analyses (as opposed to validity analyses where statistical significance is the primary concern), as long as the sample is of sufficient size to provide a stable estimate of subgroup means and standard deviations.

The smallest subgroup in this study (blacks) has a sample size of 50. For the final predictor, a sample of 50 is sufficient to place a 95% confidence band of width 6.7 points about the group mean. In other words, we can be 95% confident that the true group mean for blacks lies within $\pm$ 3.3 points of the obtained mean score of 133.7.

For the largest subgroup (whites) the corresponding 95% confidence interval spans 3.2 points. Thus, while the 95% confidence interval for the smallest subgroup is slightly more than twice that of the largest sub group, it is small relative to the total range of scores (110 points) and therefore provides a quite accurate estimate of the true subgroup mean.

Estimating the adverse impact of the predictor upon subgroups (a major fairness analysis of this project) requires accurate estimates of subgroup mean scores. The sample sizes in this project are large enough to provide the needed accurate estimates and thus are technically adequate for the fairness analyses conducted.

Normative Data. Normative data must provide an accurate estimate of the performance of incumbents in the validation sample on the Corrections Officer Examination. As was discussed, even the smallest subgroup in this project is large enough to provide quite accurate estimates of subgroup mean scores. Predictor data are available for a total of 339 corrections officer incumbents. Using these data we can be 95% confident that the true average score of corrections officers lies within $\pm$ 1.4 points of the obtained mean of 134.4. There can be little doubt as to the expected performance of corrections officer incumbents on the Corrections Officer Examination thereby establishing the technical adequacy of the final validation sample for use in determining test norms.

Summary

The original project plan called for the testing of 500 corrections officer incumbents. Due to factors beyond the control of PDI or STC only 339 incumbents were tested. Of these 339 only 253 incumbents had their job performance rated. As a result, fairness and validation analyses were conducted using fewer data than was originally planned.

The adequacy of statistical tests for a given sample is a joint function of sample size and the expected magnitude of the statistic (e.g., validity coefficient, group mean) being tested. The power of the three major analyses performed in developing and valdating the Corrections Officer Examination has been described above. In all cases, the tests had sufficient power to detect statistics of the magnitude expected in this project. Thus the validation sample provides a firm basis for evaluating the appropriateness of using the Corrections Officer Examination to screen job candidates.

In the future STC should continue to collect gender, ethnic status, and examination score data in order to monitor the validity and fairness of the Corrections Officer Examination and to augment the norms upon which selection decisions will be made. Further recommendations concerning the use and monitoring of the Corrections Officer Examination can be found in the Corrections Officer Validation Report and related documents available from the Board of Corrections.

References

Boehm, V.R. Are we validating more but publishing less? Personnel Psychology, 1982, 35, 175-187.

Schmidt, F.L., Hunter, J.F., & Urry, V.W. Statistical Power in criterion-related validation studies. Journal of Applied Psychology, 1976, 61, 473-485.

Figure 1. Statistical power of a t test with 277 df as a function of effect magnitude.

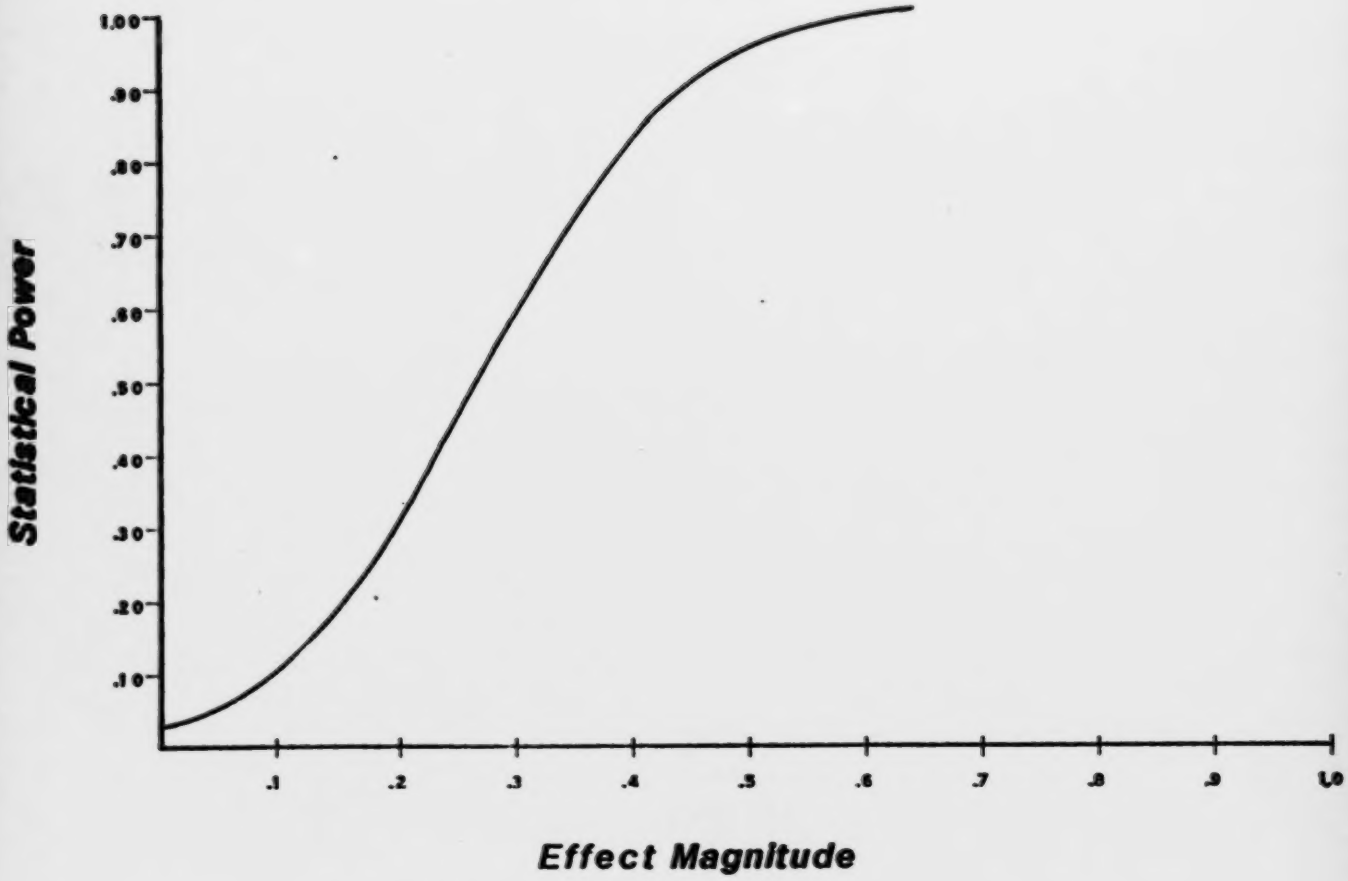
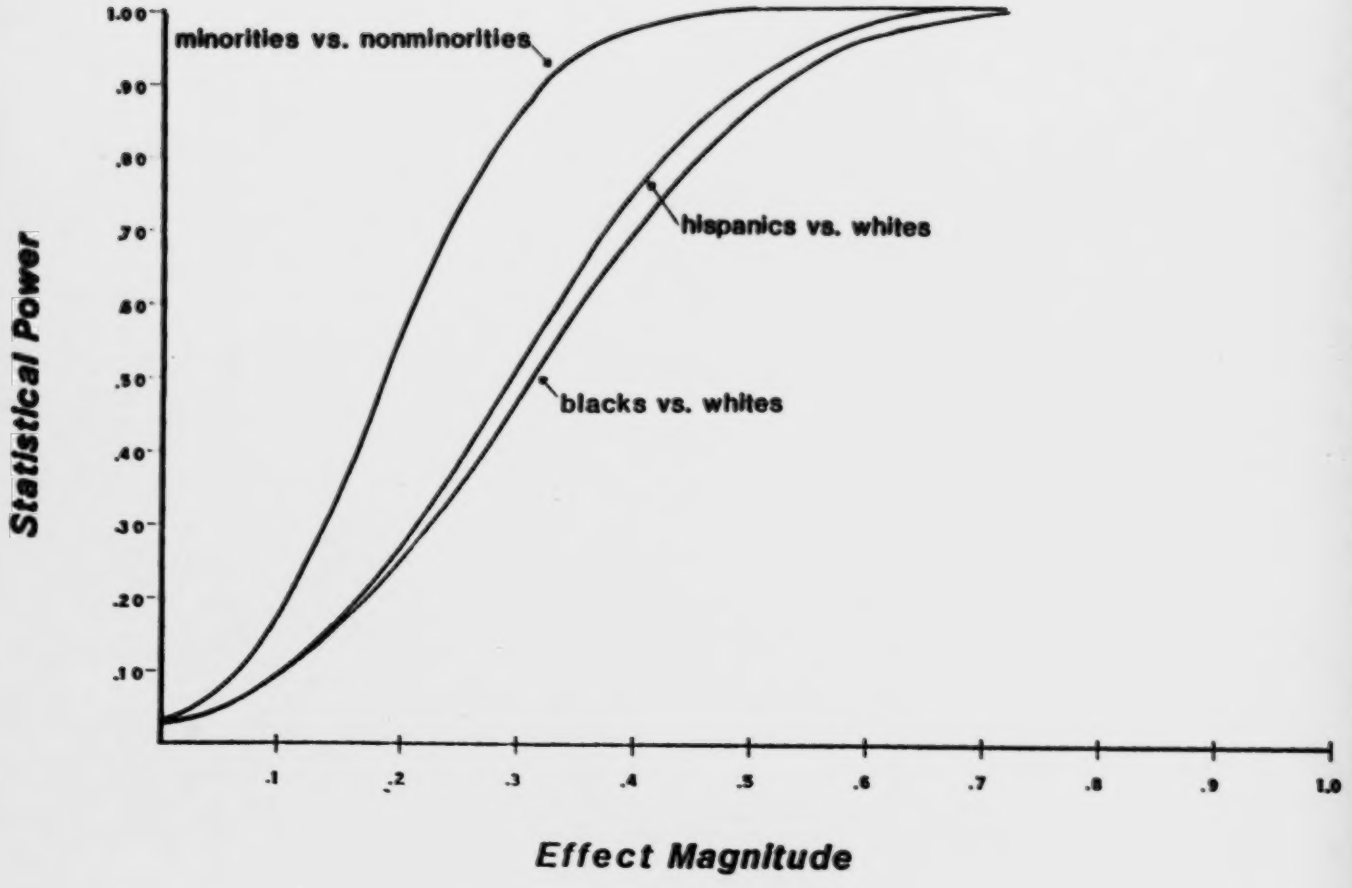


Figure 2. Statistical power of t tests with 265, 215, and 205 df as a function of effect magnitude.



Technical Appendix

Below is a brief outline of the computations used to determine the power of a sample of 253 to detect a correlation coefficient of .22 or greater.

Schmidt, Hunter, and Urry (1976) give the following equation for use in determining the sample size necessary to achieve a established test with specified power.

$$N = \frac{(Z_1 - Z_2)^2}{E(F_z)^2} + 3$$

Where: Z_1 = z score corresponding the alpha level of the statistical test (alpha is typically set at .05);

Z_2 = z score corresponding to the power of the statistical test;

N = The sample size necessary to achieve a given power for a specified alpha level; and

$E(F_z)$ = the expected value of the test statistic.

Solving the equation to isolate Z_2 yields:

$$Z_2 = Z_1 - E(F_z)\sqrt{N-3}.$$

For corrections officers we want to know the power of a statistical test based upon N = 253 and $E(F_z)$ = .220 at an alpha level of .05 (Z_1 = 1.96). Substituting into the above equation gives Z_2 = 1.52. Using a table of the normal distribution it is found that a z-score of 1.52 corresponds to a power of .94.